1

Technical Report 757

Project A

Improving the Selection, Classification and
Utilization of Army Enlisted Personnel

Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS

Robert H. Davis, Gregory A. Davis,
John N. Joyner, and Maria Veronica de Vera
Human Resources Research Organization

Selection and Classification Technical Area

Manpower and Personnel Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

August 1987

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON *Technical Director*	**WM. DARRYL HENDERSON** **COL, IN** *Commanding*

Research accomplished under contract
to the Department of the Army

Technical review by

Clessen Martin
Clint Walker

QUALITY INSPECTED 2

Accesion For	
NTIS CRA&I	☑
DTIC TAB	□
Unannounced	□
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

NOTICES

~~DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.~~

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 757

6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization	6b OFFICE SYMBOL *(If applicable)* HumRRO	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS *(City, State, and ZIP Code)* 1100 South Washington Street Alexandria, Virginia 22314-4499		7b. ADDRESS *(City, State, and ZIP Code)* 5001 Eisenhower Avenue Alexandria, Virginia 22333-5600
8a. NAME OF FUNDING/SPONSORING ORGANIZATION --	8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA 903-82-C-0531
8c. ADDRESS *(City, State, and ZIP Code)* --		10. SOURCE OF FUNDING NUMBERS

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
	2Q263731A792		

11. TITLE *(Include Security Classification)*
Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS

12. PERSONAL AUTHOR(S)
Robert H. Davis, Gregory A. Davis, John N. Joyner, and Maria Veronica de Vera (HumRRO)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
Final Report	FROM Oct 1983 TO Sep 1985	August 1987	72

16 SUPPLEMENTARY NOTATION Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute).

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Classification, Criterion measures, Knowledge tests, MOS-specific tests, Predictor measures, Project A field test, Selection, Soldier effectiveness, Training performance

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The research described in this report was performed under Project A, the U.S. Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This research sought to develop tests that will provide information about the performance of soldiers in training. Specifically, this task was (1) to create reliable and content-valid Job-Relevant Knowledge Tests (JRKTs) for 19 Military Occupational Specialties (MOS) that can measure the cognitive component of training success, and (2) to develop the JRKTs to predict first- and second-tour job performance.

This report describes the methods used to develop the 19 JRKTs, and the characteristics of the various test versions as they evolved from the initial item pools. The JRKTs were developed in three batches (A, B, and Z) consisting of 4, 5, and 10 MOS, respectively. Initial item pools were based on Army Occupational Survey Programs, Programs of Instruction,
(continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL Lawrence M. Hanser	22b. TELEPHONE *(Include Area Code)* (202) 274-8275	22c. OFFICE SYMBOL PERI-RS

83 APR edition may be used until exhausted.
All other editions are obsolete.

ARI Technical Report 757

19. Abstract (continued)

and other relevant Army reference materials. Job incumbents and school trainers reviewed the test items for technical accuracy and for importance and relevance to Skill Level 1 soldiers. The test items were also administered to groups of trainees in their last week of training. Batches A and B were field tested with job incumbents, and the test item parameters were analyzed; Batch Z will, in effect, be field tested in the Concurrent Validation test administration (June-November, 1985). Based on the available data, each JRKT was carefully tailored to ensure that the test content is a reliable and valid representation of training success.

The appendixes that provide further documentation for this research are contained in a separate report, with limited distribution:

Appendixes to ARI Technical Report 757: Develonment and Field Test of Job-Relevant Knowledge Tests for Selected MOS (ARI Research Note in preparation).

Volume 1: Appendix A, Guide for Test Administrator
Appendix B (Part 1), Job-Relevant Knowledge Tests for MOS 11B, 12B, 13B, 16S, 19E, 27E, 31C, 51B, 54E

Volume 2: Appendix B (Part 2), Job-Relevant Knowledge Tests for MOS 55B, 63B, 64C, 67N, 71L, 76W, 76Y, 91A, 94B, 95B

[illegible]

Technical Report 757

Project A
Improving the Selection, Classification and
Utilization of Army Enlisted Personnel

Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS

Robert H. Davis, Gregory A. Davis,
John N. Joyner, and Maria Veronica de Vera
Human Resources Research Organization

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

August 1987

Army Project Number
2Q263731A792

Manpower and Personnel

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This document describes the development and field testing of job-relevant knowledge tests for evaluating the training performance of enlisted personnel. The research was part of Project A, the Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance.

Project A is being conducted under contract to the Selection and Classification Technical Area (SCTA) of the Manpower and Personnel Research Laboratory (MPRL) at the U.S. Army Research Institute for the Behavioral and Social Sciences. The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A." This research supports the MPRL and SCTA mission to improve the Army's capability to select and classify its applicants for enlistment or reenlistment by ensuring that fair and valid measures are developed for evaluating applicant potential based on expected job performance and utility to the Army.

Project A was authorized through a Letter, Deputy Chief of Staff for Operations and Plans (DCSOPS), "Army Research Project to Validate the Predictive Value of the Armed Services Vocational Aptitude Battery," effective 19 November 1980; and a Memorandum, Assistant Secretary of Defense (MRA&L), "Enlistment Standards," effective 11 September 1980.

In order to ensure that Project A research achieves its full scientific potential and will be maximally useful to the Army, a governance advisory group comprised of Army General Officers, Interservice Scientists, and experts in personnel measurement, selection, and classification was established. Members of the latter component provide guidance on technical aspects of the research, while general officer and interservice components oversee the entire research effort; provide military judgment; provide periodic reviews of research progress, results, and plans; and coordinate within their commands. Members of the General Officers' Advisory Group include MG Porter (DMPM) (Chair), MG Briggs (FORSCOM, DCSPER), MG Knudson (DCSOPS), BG Franks (USAREUR, ADCSOPS), and MG Edmonds (TRADOC, DCS-T). The General Officer's Advisory Group was briefed in May 1985 on the issue of obtaining proponent concurrence of the criterion measures before administering the concurrent validation. Members of Project A's Scientific Advisory Group (SAG), who guide the technical quality of the research, include Drs. Milton Hakel (Chair), Philip Bobko, Thomas Cook, Lloyd Humphreys, Robert Linn, Mary Tenopyr, and Jay Uhlaner. The SAG was briefed in October 1984 on the results of the Batch A field test administration. Further, the SAG was briefed in March 1985 on the contents of the proposed Trial Battery.

FOREWORD (Continued)

A comprehensive set of new selection/classification tests and job performance/training criteria have been developed and field tested. Results from the Project A field tests and subsequent concurrent validation will be used to link enlistment standards to required job performance standards and to more accurately assign soldiers to Army jobs.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

Mike Rumsey, Rebecca Oxford-Carpenter, Clessen Martin, and Clint Walker of the U.S. Army Research Institute (ARI) provided invaluable assistance in many ways during the course of development of the Job-Relevant Knowledge Tests (JRKTs). John Campbell of the Human Resources Research Organization (HumRRO) has kept our vision clearly focused on the scientific tasks at hand and lent us some of his psychometric prowess as well. Kent Eaton of ARI has provided us the opportunity to be clear and precise in our work. Laurie Wise and Winnie Young of American Institutes for Research provided excellent and timely expertise and support in data analysis at various points in this effort. Robert McKenzie of HumRRO was responsible for developing the Batch Z JRKT and modifying much of the Batch A JRKTs.

The greatest and most heart-felt acknowledgment goes, however, to all the Army personnel who contributed their time and effort to the development, pilot testing, and field testing of the measures of the JRKTs. They are too numerous to mention, but we thank them all for the sincere effort and high motivation they showed as they worked with us on Project A.

EXECUTIVE SUMMARY

Requirements:

The general purpose of the Project A research on training criteria is to generate information about training performance to validate initial predictors and to predict first-tour and second-tour performance in the Army.

The overall goal of Task 3 of Project A is to develop tests that will provide information about the performance of soldiers in training. Specifically, the main objectives are as follows:

1. To create reliable and content-valid Job-Relevant Knowledge Tests (JRKTs) for 19 Military Occupational Specialties (MOS) that can measure the cognitive component of training success.

2. To develop the JRKTs to predict first- and second-tour job performance.

Procedure:

The JRKTs were developed in three batches (A, B, and Z) consisting of 4, 5, and 10 MOS, respectively. Development took place from October 1983 to May 1985.

The steps in the construction of the JRKTs were as follows:

1. Development of initial item pool
2. Review by job incumbents
3. Review by school trainers
4. School test administration
5. Preparation for field test of Batches A and B MOS
6. Field test with job incumbent
7. Review by Training and Doctrine Command (TRADOC) proponent
8. Preparation for Concurrent Validation

The initial item pool was written by Project A research staff. The Army Occupational Survey Programs, Programs of Instruction, Soldier Manuals, and other pertinent Army reference manuals were used in drafting the test items.

Job incumbents, serving as subject matter experts (SMEs), reviewed the test items for technical accuracy and appropriate vocabulary, and rated item content for importance and relevance to Skill Level 1 soldiers (judged in three scenarios--combat, combat readiness, and garrison duty). Similarly, items were reviewed by school trainers and rated for their importance in training. Test items were then administered to groups of trainees in their last week of training. After items were revised in accordance with comments from the various reviews, the item pools were prepared for field test administration to job incumbents.

Field testing was conducted in two phases--from March through September 1984 for the Batch A MOS, and from February through April 1985 for the Batch B MOS. For the other 10 MOS, known collectively as Batch Z, the next major data collection, the Concurrent Validation (CV), will be the *de facto* field test. The methods used to develop the three batches (A, B, and Z) differed very little, and only insofar as experience in the development of each batch inspired improvements in procedure for ensuing development work. Review by the Proponent agencies for the individual MOS preceded preparation of the tests for CV administration.

Findings:

The effort to create content-valid and reliable Job-Relevant Knowledge Tests for measuring the cognitive components of training success can be evaluated against three criteria of content validity: domain clarity, content representativeness, and content relevance.

First, the domain for each MOS was operationally identified and items were drawn from that domain on the basis of item budgets. With respect to the second criterion, content representativeness, the proportions of items assigned to different duty areas on different versions of the tests were similar and reflected areas of the MOS judged to be important and relevant. In a few cases, it was found that some duty areas were no longer performed as a part of an MOS or that an MOS had been given some new responsibility, but changes of this magnitude were rare. With respect to the third criterion, content relevance, the elaborate procedure for determining relevance addressed this need. Items judged as not relevant to the job were eliminated; moreover, relevance was judged in terms of importance, with only those items judged to be very important on one or more of the three scenarios retained. Every effort was made, when items were reviewed by subject matter experts, to ensure that the review groups were balanced for race and gender.

The tests can also be evaluated in terms of more traditional psychometric properties, particularly reliability. All of the tests had relatively high reliability coefficients. Alpha of tests administered to job incumbents ranged from .76 for MOS 95B to .93 for MOS 19E, with a mean reliability across all nine tests of .88.

Utilization of Findings:

Based on the data presented, one can conclude that the JRKT versions developed are reliable and content-valid measures of the cognitive component of training success. The test evaluations of the SMEs and the field test analyses were considered in preparing the JRKTs for Concurrent Validation. All pre-Concurrent Validation JRKT versions were then submitted to the appropriate TRADOC Proponent for review. The Proponent evaluated and updated the test, and deleted, modified, or added items as appropriate. Based on the available data, each JRKT was carefully tailored to ensure that the test content was a reliable and valid representation of training success, suitable for use in the Concurrent Validation.

DEVELOPMENT AND FIELD TEST OF JOB-RELEVANT KNOWLEDGE TESTS FOR SELECTED MOS

CONTENTS

Page

Chapter

LIST OF TABLES

*NOTE: The Appendixes to this report are contained in a separate report with limited distribution:

Appendixes to ARI Technical Report 757: Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS (ARI Research Note in preparation).

Volume 1: Appendix A, Guide for Test Administrator Appendix B (Part 1), Job-Relevant Knowledge Tests for MOS 11B, 12B, 13B, 16S, 19E, 27E, 31C, 51B, 54E

Volume 2: Appendix B (Part 2), Job-Relevant Knowledge Tests for MOS 55B, 63B, 64C, 67N, 71L, 76W, 76Y, 91A, 94B, 95B

Page

LIST OF FIGURES

DEVELOPMENT AND FIELD TEST OF JOB-RELEVANT KNOWLEDGE TESTS FOR SELECTED MOS

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program which the U.S. Army has undertaken to develop an improved personnel selection and classification system for the enlisted ranks. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through the use of new and improved selection/classification tests which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.

- Develop and validate new selection and classification measures.

- Validate intermediate criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), so that better informed reassignment and promotion decisions can be made throughout a soldier's career.

- Determine the relative utility to the Army of different performance levels across MOS.

- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The research design for the project incorporates three main stages of data collection and analyses in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from Army accessions in fiscal years (FY)

1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB) and their subsequent performance in training and their scores on the first-tour Skills Qualification Tests (SQT).

In the second iteration, a concurrent validation design will be executed with FY83/84 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/ personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS). The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered
perceptual and psychomotor measures, will be administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

In the third iteration (the Longitudinal Validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3,500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

For both the concurrent and longitudinal validations, the sample of MOS was specially selected as representative of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45% of Army accessions. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

Activities and progress during the first 3 years of the project were reported as follows: for FY83, in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37; for FY84, in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 84-14; for FY85, in ARI Technical Report 746 and an ARI Research Note (in preparation). Other publications on specific activities during those years are listed in those annual reports.

For administrative purposes, Project A is divided into five research tasks:

Task 1 -- Validity Analyses and Data Base Management
Task 2 -- Developing Predictors of Job Performance
Task 3 -- Developing Measures of School/Training Success
Task 4 -- Developing Measures of Army-Wide Performance
Task 5 -- Developing MOS-Specific Performance Measures

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries that will be used in the comprehensive Concurrent Validation program which is being initiated in FY85.

The present report is one of five that have been prepared under Tasks 2-5 to report the development of the measures and the results of the field tests, and to describe the measures to be used in Concurrent Validation. The five reports are:

Task 2 -- Development and Field Test of the Trial Battery for Project A, Norman G. Peterson, Editor, ARI Technical Report 739, May 1987.

Task 3 -- Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert H. Davis, et al., ARI Technical Report 757, August 1987.

Task 4 -- Development and Field Test of Army-wide Rating Scales and the Rater Orientation and Training Program, Elaine D. Pulakos and Walter C. Borman, Editors, ARI Technical Report 716, July 1986.

Task 5 -- Development and Field Test of Task-Based MOS-Specific Criterion Measures, by Charlotte H. Campbell, et al., ARI Technical Report 717, July 1986.

-- Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, by Jody L. Toquam, et al., ARI Technical Report (in preparation).

Chapter 1

THE OBJECTIVES

The general purpose of the Project A research on training criteria is to generate information about training performance that can be used in the validation of initial predictors and in the prediction of first-tour and second-tour performance in the Army.

To accomplish this purpose, tests that measure training success have been developed. As job performance surrogates, training measures can serve to reduce the time required to validate predictors from years to months. When used to predict subsequent performance, training measures can increase the accuracy of MOS classification over that obtained with preinduction predictors alone. Both the extent to which training measures can be used as surrogates for ultimate job performance criteria, and the degree of incremental validity obtained by including training success itself as a predictor, will be assessed during the course of Project A.

The overall goal of Task 3 is to develop tests that will provide information about the performance of soldiers in training. Specifically, Task 3 has two main objectives:

(1) To create reliable and content-valid Job-Relevant Knowledge Tests (JRKT) for 19 Military Occupational Specialties (MOS) that can measure the cognitive component of training success.

(2) To develop the JRKT to predict first- and second-tour job performance.

This report describes the methods used to develop the JRKT for the 19 MOS, and the characteristics of the various test versions as they evolved from the initial item pools. A 20th MOS, 19K, being developed in the summer of 1985, was included in a few of the analyses of this report, although it is not part of the Concurrent Validation. All JRKTs were pilot tested on trainees at the end of their Advanced Individual Training (AIT), and nine of the JRKTs were field tested on job incumbents.

The nine JRKTs that were field tested are referred to as the Batch A (four) and the Batch B (five) MOS. For the other 10 MOS, known collectively as Batch Z, the Concurrent Validation (CV) will be the de facto field test. The methods used to develop the three batches (A, B, and Z) differed very little, and only insofar as experience in the development of each batch inspired improvements in procedures for ensuing development work.

Chapter 2

THE MODELS

MEASUREMENT MODEL

The Construct of Training Success

As stated in the discussion of objectives, the JRKTs will be used primarily as criterion measures of the cognitive component of training success. What precisely do we mean by this phrase? As used in Project A, the term training success refers to the impact of training on individuals, not to the impact on groups or to the overall success of the program. The Project A Research Plan defines training success in terms of the individual trainee's achievement; the original Statement of Work used the term in a similar way, that is, to refer to specific measures taken on soldiers in the course of training, such as those included in the Army Training and Doctrine Command (TRADOC) Educational Data System (TREDS) and the Automated Instruction Management System (AIMS). Many of these instruments include both "hands-on" and cognitive measures.

The construct of training success, as used in Project A, encompasses the outcomes of both formal training and organizational socialization. Organizational socialization is defined as the way in which soldiers accommodate to their role as soldiers and "learn the ropes," such as the attitudes, standards, and patterns of behavior expected of soldiers in general and of soldiers in an assigned MOS. Organizational socialization is achieved through formal training, of course, but it is also developed outside of the regular classroom through a variety of activities, including role modeling, drill, stressful experiences, behavior reinforcement, and similar practices designed to produce appropriate military attitudes, social interactions, and automaticity. Furthermore, it is reasonable to suppose that a great deal of organizational socialization takes place in AIT both inside and outside of the classroom.

A wide variety of potentially useful measures either are available or could be created to assess these three major aspects of training success: (1) the cognitive component, (2) the hands-on component, and (3) the organizational socialization component. The JRKTs are designed to measure the cognitive component of formal training experiences, specifically AIT. Thus, JRKTs measure only one part of the total domain encompassed by the construct of "training success."

The cognitive component of training success includes two types of knowledge: (1) about the job as taught in AIT and (2) about a wide range of "common skills" that cut across all MOS and that all soldiers are expected to know.

Relationships Between Training and the Job

Within the military, there is a very close relationship between training content and tasks performed on the job. Skill Level 1 soldiers within any given MOS may have quite different jobs--that is, jobs that emphasize different skills--but it is almost always the case that the skills necessary for the performance of a job at Skill Level 1 are taught in AIT. As a matter of doctrine, training must be job-related, and in the development of training objectives and materials every effort is made to ensure that they are job-related. As a result, if a content-valid test is created on the basis of curricular materials alone, one can assume that most of the items will be job-related. School curricula sometimes include topics or tasks that are unrelated to the job, but this is the exception rather than the rule.

Classes of Items

As might be expected, some trainees learn important job skills that are not taught in the schools. As a result of extracurricular activities, outside study, generalization, or all three, a trainee may develop some job skills in the school setting that are not taught as part of the curriculum. From the perspective of criterion development, one might hypothesize that the exceptional--that is, most successful--trainee is one who goes beyond the formal curriculum and learns such skills.

Similarly, military training performance is predictive of later military job performance because (1) training performance reflects general learning ability (and hence identifies who will acquire knowledge on the job), (2) the information acquired in training is in itself a significant factor in job performance, or more likely (3) both.

Accordingly, two subsets of test items were constructed in this research--one reflecting training requirements, and the other job requirements. Where a sufficient number of test items could be developed for both classes, scores on the two types of items may shed light on the relationships among predictors, success in training, and success on the job. Four classes of items resulted: those relevant only to training, those relevant to both job and training, those relevant only to the job, and common items that cut across all MOS. Common items were written focusing on common soldier skills as defined by the Common Task Manual.

Emphasis on Content Validity

There is little agreement among psychologists regarding the use of the term content validity, but there is general agreement that content considerations are fundamental to all psychological measurement and that they are especially relevant to tests purporting to measure training and educational success. Although definitions of content validity differ, the literature stresses three critical components: clarity of the content domain, representativeness of content, and relevance of content.

Domain Clarity. By domain clarity we mean that the content domain should be defined unambiguously. Essentially, this means that the boundaries that outline the content domain clearly specify the subject/duty areas that define training success. At the outset of the test development process, the content domain was defined operationally by the following:

- Training: Programs of Instruction (POIs), lesson plans, technical publications, Soldier's Manuals, and Common Task Manual.

o Job: Army Occupational Surveys (AOSPs), technical publications, Soldier's Manuals, and Common Task Manual.

Content Representativeness. The issue of content representativeness refers to the question of whether or not the domain has been adequately sampled. Specifically, it involves determining whether the proportions of items allocated to the different duty areas reflect the relative importance of each duty area in relation to the entire content domain.

Operationally, establishing content representativeness involves a strategy for arriving at item budgets, that is, allocating items to areas of the content domain. When people disagree about such matters, the question is normally resolved on the basis of the level of expertise of those making the decision. In the case of the JRKTs, the strategy for developing item budgets was defined by test construction experts, but the strategy for weighting the budgets employed data from subject matter experts (job incumbents and trainers). The actual operations in this process are described in the Chapter on the development process.

Content Relevance. The issue of content relevance concerns the relevance of the content to the purpose of measurement. In the broadest sense, this issue hangs upon the purposes of Project A itself, which are discussed elsewhere, and the relevance of content domain to those purposes. But in a somewhat narrower sense, we may simply ask whether specific items are relevant to the two facets of the content domain that we have already identified, that is, training and the job. Furthermore, this question may be extended to explore the relevance of items under different circumstances or scenarios, such as peacetime, readiness, and combat. How this was accomplished is described in the sections dealing with the review of items by job incumbents and school trainers. The question of who is best qualified to make such judgments deserves some preliminary discussion, however.

Subject matter experts (SMEs) were called upon to make judgments about relevance and importance. Some people, however, are more expert than others about some parts of the domain. Officers, for example, have a different perspective from enlisted personnel, and officers, or enlisted personnel, or both may differ among themselves. Furthermore, the number of possible perspectives on any given MOS is very large. Soldiers in a light infantry division, for example, may use entirely different weapons, vehicles, and even tools than soldiers in the same MOS in another setting. Which of these various groups have the most relevant expertise?

If it were possible to bring all of the experts together in a single room, most differences undoubtedly could be explained and resolved. But in a study of this magnitude, judgments on such complicated questions are made over a fairly long period of time by experts residing in different parts of the world, and they are dealing with what are in fact very dynamic systems, in that equipment and doctrine are in a continuous state of change.

The final arbiter in this case is the "Proponent," the agency officially designated by the Army as responsible for the MOS. Frequently, the Proponent is closer to the school than to the operational environment, often being co-located with one of the schools training the MOS.

Saying that the Proponent is the final arbiter does not mean that the Proponent determined and controlled all of the content of the JRKT. From the universe of possible content and possible items, the Proponent actually reviewed and commented on a JRKT version that had been earlier submitted for SME evaluation. Before items were submitted to Proponents, the universe had been systematically sampled, and the sample items had been meticulously reviewed by subject matter experts from the operational units. The Proponent's role in this process was primarily reactive, rather than proactive. The JRKTs submitted to Proponents for review had first been subjected to a rigorous process of content selection, item budgeting, and SME review.

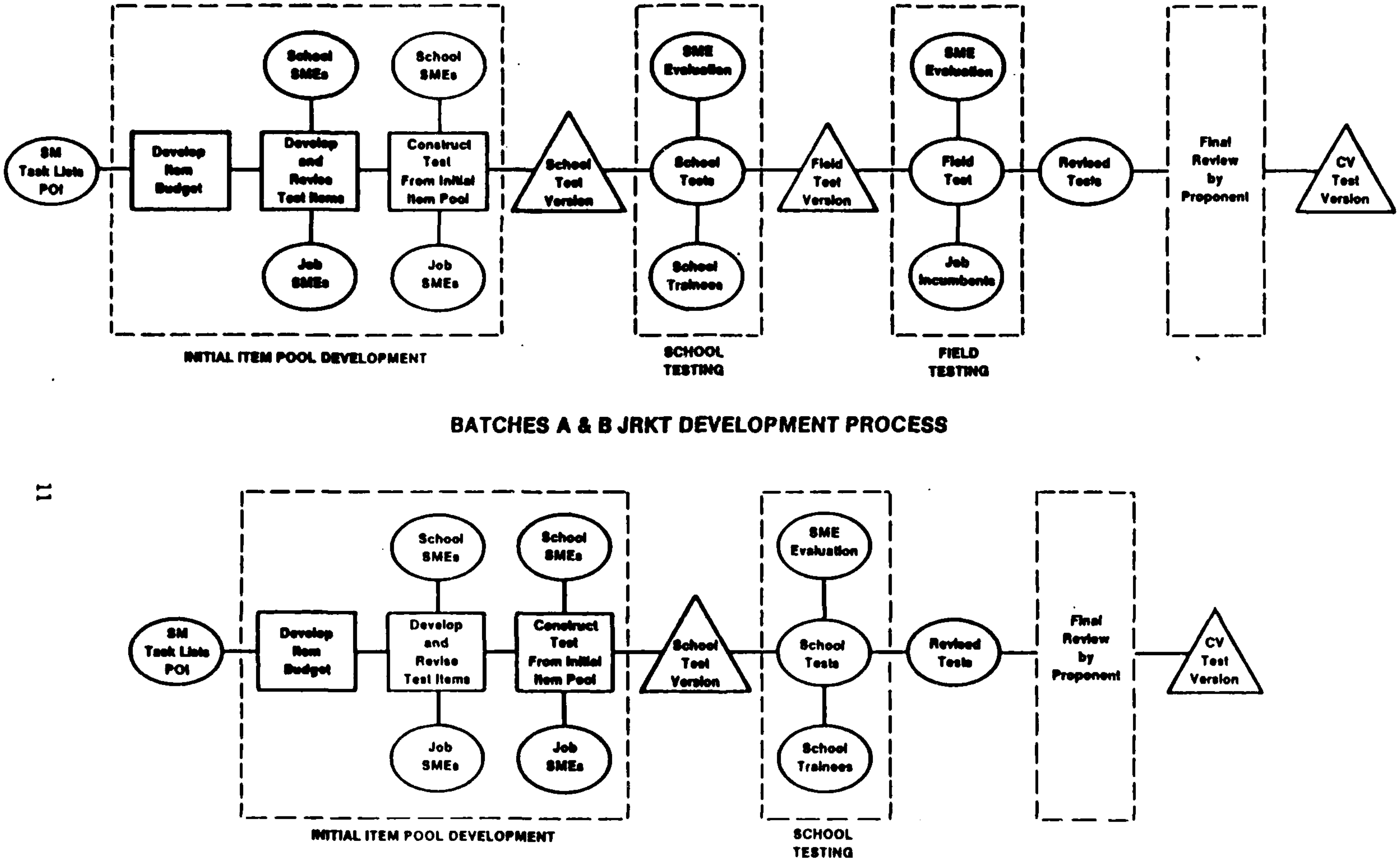
With this caveat in mind, it is nevertheless true that the final judgment about the items submitted was left to the Proponents. Proponents could accept or reject items, and suggest modifications to items or additions to a test to conform to their perception of the content domain.

One important implication of the role of the Proponents in evaluating the JRKTs is that, although the content domains had been operationalized as described above, a Proponent could, and sometimes did, introduce new considerations. But the fact of the matter is that the operationalization of the content domain was "by the book," and the Proponent was more likely to perceive generally accepted doctrine and practice as coextensive than were soldiers in the field, who frequently deviated from doctrine. On the other hand, the Proponent sometimes said that items were inappropriate for Skill Level 1 soldiers or that items were too difficult or too easy, when empirical data suggested otherwise. All such issues were discussed with the Proponents. Most were resolved without difficulty; in some cases, items in question were simply eliminated from the pool.

DEVELOPMENT MODEL

The main steps in developing the JRKTs are shown in Figure 1. The test items were reviewed by subject matter experts during the initial development/revision phase, before being administered to trainees and incumbents. Although each set of test questions went through numerous alterations as it evolved, the three main versions are: (1) the school test version, (2) the field test version, and (3) the Concurrent Validation (CV) test version. Figure 1 also summarizes the differences in developmental procedures between Batches A/B and Batch Z.

Discussion of the development of these various test versions is the subject of the following chapter, which is organized sequentially to present the developmental data for the three test versions shown in Figure 1.



BATCHES A & B JRKT DEVELOPMENT PROCESS

BATCH Z JRKT DEVELOPMENT PROCESS

Figure 1. Job-Relevant Knowledge Test (JRKT) Development Process.

Chapter 3

THE DEVELOPMENT PROCESS

The JRKTs were developed in three batches (A, B, and Z) consisting of four, five, and ten MOS, respectively (Table 1). Development took place from October 1983 to May 1985. An additional MOS, 19K, was being developed in the summer of 1985 for the Longitudinal Validation and is included in a few of the analyses in this report.

Table 1

MOS Included in Batches A, B, and Z

Batch A		Batch B	
13B	Cannon Crewman	11B	Infantryman
64C	Motor Transport Operator	19E	Armor Crewman
71L	Administrative Specialist	31C	Radio Teletype Operator
95B	Military Police	63B	Light Wheel Vehicle Mechanic
		91A	Medical Specialist

Batch Z[a]	
12B	Combat Engineer
16S	MANPADS Crewman
27E	Tow/Dragon Repairer
51B	Carpentry/Masonry Specialist
54E	NBC Specialist
55B	Ammunition Specialist
67N	Utility Helicopter Repairer
76W	Petroleum Supply Specialist
76Y	Unit Supply Specialist
94B	Food Service Specialist
19K	M1 Abrams Armor Crewman[b]

[a] Not field tested with job incumbents.

[b] Developed for Longitudinal Validation; not included in the Concurrent Validation.

As noted previously, all three JRKT batches were pilot tested at the appropriate MOS school training sites, but only Batches A and B were field tested with job incumbents (Figure 1). The Concurrent Validation will serve as the field test for job incumbents for Batch Z.

Procedures were modified somewhat on the basis of experience as the tests were developed. For example, all item pools were reviewed by groups of SMEs as described below. However, after the first few group reviews, it was apparent that a preliminary review by one SME for accuracy, correct use of technical language, currency, and appropriateness could greatly facilitate the group review. Accordingly, this step was introduced in the process, and it did indeed appear to expedite the group reviews.

Project-wide decisions also led to some modifications in the original design of the item development process. For example, a concern for racial and gender balance within SME groups reviewing items later led to the development and implementation of guidelines for taking racial and gender aspects into account in assigning SMEs to review groups. A second informal review was scheduled for all items that had been reviewed before the implementation of the guidelines. The characteristics of all SMEs who participated in the formal review are summarized in the section describing the review by job incumbents. With these few exceptions, the procedures for developing the tests were essentially the same for the various MOS.

The steps in the construction of the JRKTs, each of which will be described in greater detail below, were as follows:

1. Development of initial item pool
2. Review by job incumbents
3. Review by school trainers
4. School test administration
5. Preparation for field test of Batches A and B MOS
6. Field test with job incumbents
7. Review by TRADOC proponent agencies
8. Preparation for Concurrent Validation.

INITIAL ITEM POOL DEVELOPMENT

Development of the item pools proceeded in four steps: (1) refine the Army Occupational Survey Program (AOSP) task list, (2) calculate item budget, (3) draft items, and (4) develop the pool of items.

Refinement of AOSP Task List

The AOSP collects and analyzes data on tasks being performed by soldiers in different MOS. Within each MOS tasks are grouped into duty areas. The number of duty areas in the 19 MOS ranged from 15 to 23 (Table 2). One of the key statistics reported with respect to these duty areas, tasks, and subtasks is percentage of soldiers at different skill levels performing the task activity. As described in more detail below, this statistic was used to prepare a test item budget prior to drafting items.

Before the AOSP reports were used, however, several actions were taken to refine these data. Refinement was needed because of their publication dates (Table 3). The SME reviews provided useful information for the MOS whose AOSP publication dates were not recent (e.g., 91A - 1976).

Table 2

Illustrative List of Duty Areas for a Single MOS (11B)

AOS DUTY AREA	DESCRIPTION
A	Cannon Equipment Emplacement/Displacement
B	Firing Btry Operations During Firing
C	Firing Btry Tactical Operation Training
D	Firing Btry Section Planning
E	Firing Btry Section Training
F	General Tactical Operational Training
G	Unit Defense Training
H	FA Weapon System Operator Maintenance
I	FA Weapon Movement/Transport
J	Tracked Cargo Carrier Operations and Maintenance
K	Wheeled Vehicle Operations and Maintenance
L	Preventive Maintenance Operations
M	FA Weapons Organizational Maintenance
N	Individual Weapons Training
O	Crew Served Weapons Training
P	Physical Security
Q	Ammunition Handling and Maintenance
R	Personnel Supervision
S	Land Navigation/Map Reading
T	Recon/Security/Combat Patrol Training
V	Communications Equipment and Operator Maintenance

Table 3

Publication Dates for the Different AOSP Task Lists

Batch A		Batch B		Batch Z	
MOS	Year	MOS	Year	MOS	Year
13B	1982	11B	1981	12B	1978
64C	1982	19E	1982	16S	1982
71L	1982	31C	1980	27E	1979
95B	1982	63B	1977	51B	1981
		91A	1976	54E	1981
				55B	1983
				67N	1978
				76W	1978
				76Y	1983
				94B	1983

For Batches A and B, the AOSP listings were cut as follows:

Ninety-nine percent confidence intervals were computed on the mean percentage performing all tasks. This confidence interval was calculated using the formula 2.5 pq/n, where p is the average (taken over all tasks) of the percent performing at Skill Level 1, q is 1-p, and n is the number of Skill Level 1 soldiers in the survey. Tasks with a very low percentage performing (equal to or less than the lower bound of the confidence interval) were deleted from consideration.

The remaining task statements were reformatted and then reviewed by SMEs. The purposes of this review were to:

(1) Delete AOSP statements for any of three reasons: They were no longer part of the job due to changes in doctrine or equipment; they were not really tasks, and should not have been included in the AOSP listing (e.g., administrative labels that had been misconstrued as tasks); or they were sets of tasks (i.e., they contained only individual tasks that were already in the domain).

(2) Confirm the grouping of AOSP tasks under duty areas.

For Batch Z, SME reviewers evaluated all tasks and subtasks on the AOSP.

Calculation of Item Budgets

To ensure that the content of item pools was representative of tasks performed and that it covered the entire MOS rather than aspects easiest to write items about, an item budget was drafted based on the duty areas into which the AOSP survey is divided. As previously noted there are 15 to 23 duty areas in the 19 AOSP surveys analyzed. It was expected that during tryout, revision, and field testing, items would be eliminated from the pool because of faulty construction or lack of discriminatory or predictive power. To allow for item attrition, the initial target was 225 draft items for each MOS, even though the final version of the test was expected to be closer to 150 items. AOSP data on percentage performing were used in building the budget as described below.

Determine the Match Between AOSP Duty Areas and Training Objectives. A matrix (e.g., Figure 2) was prepared to display the duty areas of the AOSP versus the subdivisions of the Program of Instruction (POI), each of which covers a number of training objectives. (In some courses, an "objective" is a major subdivision of content, but the term usually denotes small units of training.) When the AOSP duty areas were compared to training lessons by means of the matrix, three outcomes were possible: (1) some duty areas and training lessons matched completely; (2) some duty areas did not match any training lesson; (3) some training lessons did not match any duty area.

The majority of the first 200 items in the item budget were allocated to the first two categories. Combined, they constitute the job performance domain defined by the AOSP (including the intersection of the job performance domain with the training performance domain defined by the POI). Approximately 25 additional items were thus allocated to the third category, the subdivisions of the training course that had no counterparts among the duty areas of the AOSP. This category was expected to be small because of the Army's efforts to make training job-relevant.

Distribute the First 200 Items. The next activity in establishing a budget was to determine a target number of items for each duty area. The 200 items budgeted to the job performance domain were distributed across the duty areas in proportion to the mean percentage of members reported by the AOSP as performing the tasks that composed the duty area.

Within each of the AOSP duty areas, items were budgeted in proportion to how much they were emphasized in training: the greater the overlap between the AOSP tasks (within a duty area) and the training objectives (within the POI), the more items were written to represent job/training content. Figure 2 illustrates how the AOSP/POI matrix was used to calculate the degree of overlap. The formula used for this purpose was:

$$\frac{\text{Number of tasks that overlap in the training and duty areas}}{\text{Total number of tasks in duty areas}} \times \text{Number of items budgeted to duty area} = \text{Number of job/training items}$$

The remaining items (out of the original 200) were assigned to job-only content. For example, if 20 items were assigned to a duty area that had a total of eight tasks, six of which matched POI objectives, then 15 training/job items (6/8 X 20 = 15) and 5 job-only items would be written for the duty area (15 + 5 = 20).

Distribute the Remaining Items. The remainder of the item budget for a given MOS was reserved for items not related to any area of the AOSP task list, but covering training content as defined by the POI. These are

Program of Instruction (POI)	Army Occupational Survey Program (AOSP)					
	Duty Area B I. Operating Vehicle Vehicle Driving	Duty Area E Loading/Transporting/ Unloading Cargo and Equipment	Duty Area H Vehicle Operator Administration	Duty Area Q Vehicle Inspection	Hours of Instruction	Number of Items for Objectives Not in AOSP
Lesson A01 Make Operator Entries in Forms and Records			X	X	6	
Lesson B03 Drive 1/4 Ton Vehicle With Manual Transmission	X				10	
Lesson B05 Park Vehicle Parallel					1	1
Lesson C21 Transport Dangerous and Hazardous Cargo		X			2	
Number of Tasks that Overlap in the Training & Duty Areas	27	21	11	3		
Total Number of Tasks in Duty Areas	27	30	29	27		
Number of Test Items Budgeted to Duty Areas	20	14	12	4		
Number of Job/Training Items	20	10	5	0		

X - indicates an overlap between one or more AOS task or task element statements in a given duty area and one or more objectives in a lesson.

Figure 2. Matrix for developing an item budget (An abridged example from MOS 64C).

indicated in Figure 2 by an entry in the column at the right of the matrix: the number of hours of training specified for the lesson.

To calculate the number of items to be budgeted for this category, the mean number of test items already budgeted per hour of instruction was computed (the number of test items is a constant 200; the number of hours of instruction varies by MOS). The training program hours for lessons for which no AOSP match occurred was then multiplied by this number.

Thus, within the portion of the training performance domain that did not match any portion of the job performance domain, the allocation of test items was based on the amount of training time devoted to particular content.

Drafting of Items

After item budgets were established, written materials dealing with job and training activities were examined for information that could be transformed into multiple-choice test items. Five sources were used: the AOSP task lists, training materials (POIs, lesson plans, lesson guides, etc.), technical publications (Army Regulations, Technical Manuals, Field Manuals, etc.), the Soldier's Manual for each MOS, and the Common Task Manual. The Soldier's Manual is a description of the tasks that each MOS holder is to have mastered to be considered qualified at a given skill level. For developing the JRKTs, the level of interest was the entry (apprentice) level, Skill Level 1.

Development of Initial Item Pool

The initial item pool was written by Project A research staff. The item budgets were used to ensure that items were written to cover all the important Skill Level 1 tasks of the MOS. Multiple-choice items were written based on the available documents. The resulting item pools were presented to job incumbents and school trainers for their review, as described below.

REVIEW BY JOB INCUMBENTS

To prepare the item pool for review by job incumbents and school trainers, it was first reviewed by one subject matter expert, usually a senior officer. With that early input, the item pool was polished and purged of surface distractions.

The items were then reviewed by job incumbents during site visits (Table 4). On each visit, job incumbents reviewed items for technical accuracy and appropriate vocabulary, and rated item content for importance and relevance to Skill Level 1 soldiers.

Table 4

Number of Subject Matter Experts Participating in Reviews and Locations of Reviews

	Refinement of Task List		Job Incumbent Review		School Trainer Review	
MOS	No. of SME	Location	No. of SME	Location	No. of SME	Location
Batch A						
13B	5	Ft. Ord	7	Ft. Ord	7	Ft. Sill
64C	4	Ft. Ord	4	Ft. Ord	6	Ft. Dix
71L	4	Ft. Ord	6	Ft. Ord	6	Ft. Jackson
95B	5	Ft. Ord	8	Ft. Sill/Dix	10	Ft. McClellan
Batch B						
11B	5	Ft. Ord	5	Ft. Ord	6	Ft. Benning
19E	5	Ft. H. Liggett	5	Ft. H. Liggett	6	Ft. Knox
31C	5	Ft. Ord	5	Ft. Ord	6	Ft. Gordon
63B	5	Ft. Ord	5	Ft. Ord	6	Ft. Dix
91A	5	Ft. Ord	5	Ft. Ord	6	Ft. Sam Houston
BatchZ						
12B	5	Ft. Ord	6	Ft. Lewis	6	Ft. L. Wood
16S	5	Ft. Ord	5	Ft. Lewis	6	Ft. Bliss
27E	4	Ft. Ord	6	Ft. Lewis	6	Redstone Arsenal
51B	4	Ft. Ord	4	Ft. Lewis	4	Ft. L. Wood
54E	5	Ft. Ord	5	Ft. Lewis	5	Ft. McClellan
55B	5	Ft. Ord	6	Ft. Lewis	5	Redstone Arsenal
67N	5	Ft. Ord	6	Ft. Lewis	6	Ft. Rucker
76W	5	Ft. Ord	6	Ft. Ord	6	Ft. Lee
76Y	5	Ft. Ord	6	Ft. Ord	6	Ft. Lee
94B	5	Ft. Ord	8	Ft. Sill/Dix	10	Ft. McClellan

19K			7	Ft. Knox	10	Ft. Knox

Characteristics of Reviewers

Table 5 describes the characteristics of SMEs, both job incumbents and school trainers, who reviewed and rated items. For each MOS, the groups of SMEs are classified by type, rank, and race. SMEs had an average of 8.4 years of experience (SD = 2.5 years).

Table 5

Characteristics of Subject Matter Experts Reviewing and Rating Items

		Type of SME		Rank					Race				
MOS	Total No. SMEs	Job Incumbents	School Personnel	SFC	SSG	SP/5 SGT	E4 CPL or SP/4	Other	Caucasian	Black	Hispanic	Other	Mean Time in MOS
Batch A													
13B	20	7	13	5	7	4	3	1		a			10.1
64C	12	4	8	1	5	5	-	1		a			a
71L	8	2	6	1	4	3	-	-		a			6.0
95B	21	11	10	5	6	4	4	2		a			13.7
Batch B													
11B	18	5	13	3	9	-	3	3	8	4	5	1	9.2
19E	14	5	10	8	4	3	-	-	8	6	1	-	7.1
31C	18	5	13	4	9	4	-	1	9	6	1	2	9.1
63B	16	5	11	3	6	3	3	1	6	6	4	-	9.0
91A	16	5	11	8	5	-	-	3	7	7	2	-	10.8
Batch Z													
12B	12	6	6	2	2	6	-	2	6	2	4	-	7.2
16S	14	5	9	8	3	3	-	-	10	3	1	-	5.0
27E	14	6	8	-	9	5	-	-	6	6	2	-	6.0
51B	12	4	8	3	4	5	-	-	9	3	-	-	6.6
54E	16	5	11	5	5	5	-	1	9	6	1	-	8.2
55B	17	6	11	2	8	2	-	5	7	8	2	-	11.6
67N	13	6	7	-	0	4	-	-	10	2	1	-	8.7
76W	15	7	8	4	7	4	-	-	5	4	6	-	8.7
76Y	16	5	11	7	4	4	1	-		a			8.6
94B	18	6	12	8	4	4	-	2	11	5	2	-	11.3

19K	17	7	10	-	6	10	1	-	10	6	1	-	3.7
Summary Percentage	100.0	36.4	63.6	25.0	37.7	25.3	4.9	7.1	52.3	32.0	14.2	1.2	8.4

[a] No record.

To determine whether minority racial groups were underrepresented in the SME sample, a chi-square test of goodness of fit was computed from the data shown in Table 6. A comparison of the expected and observed frequencies indicates that minority groups were adequately represented in the SME reviewers.

Table 6

Distribution of Soldiers in Four Race Categories, Army-Wide and Among Subject Matter Expert Reviewers

Race	Army-Wide Percent Active Duty[a]	Expected Frequency in SME Sample	Observed Frequency in SME Sample
Caucasian	61.8	142.8	121
Black	30.5	70.4	74
Hispanic	4.0	9.2	33
Other	3.7	8.6	3
			231

[a] Source: Dr. Mark J. Eitelberg, personal communication.

Evaluation of Test Items

Item Quality. To establish the technical accuracy and appropriateness of the draft items, job incumbents were asked:

- o Would the item be clear to someone taking the test?
- o Is the keyed option really the correct answer?
- o Is there more than one correct option?
- o Are the distractors realistic and believable?
- o Is each technical term commonly used and easily understood?
- o Are there other more commonly used terms that should be included to make the question clearer?

Items were then revised on the basis of the evaluation from the incumbents (e.g., distractors were replaced by more realistic ones, stems were modified).

Importance Ratings. To establish the importance of the knowledge represented in the test items, job incumbents were asked to rate each item in the initial item pool. The ratings were of items' importance for Skill Level 1 soldiers in three different contexts: combat (Scenario 1), combat readiness (Scenario 2), and garrison duty (Scenario 3). The scenarios used to describe these three contexts are shown in Figure 3.

1) Your unit is assigned to a U.S. Corps in Europe. Hostilities have broken out and the Corps combat units are engaged. The Corps' mission is to defend, then reestablish, the host country's border. Pockets of enemy airborne/heliborne and guerilla elements are operating throughout the Corps sector area. The Corps maneuver terrain is rugged, hilly, and wooded, and weather is expected to be wet and cold. Limited initial and reactive chemical strikes have been employed but nuclear strikes have not been initiated. Air parity does exist.

2) Your unit is deployed to Europe as part of a U.S. Corps. The Corps' mission is to defend and maintain the host country's border during a period of increasing international tension. Hostilities have not broken out. The Corps maneuver terrain is rugged, hilly, and wooded, and weather is expected to be wet and cold. The enemy approximates a combined arms army and has nuclear and chemical capability. Air parity does exist. Enemy adheres to same environmental and tactical constraints as does U.S. Corps.

3) Your unit is stationed on a post in the Continental United States. The unit has personnel and equipment sufficient to make it mission capable for training and evaluation and installation support missions. The training cycle includes periodic field exercises, command and maintenance inspections, ARTEP evaluations, and individual soldier training/SQT testing. The unit participates in post installation responsibilities such as guard duty and grounds maintenance and provides personnel for ceremonies, burial details, and training support to other units.

Figure 3. Alternative scenarios for judging importance of tasks and items.

A 5-point scale was used to collect importance ratings:

1 Of little importance
2 Somewhat important
3 Moderately important
4 Quite important
5 Very important

Table 7 shows the percentage of items in the initial item pool rated at each of the five different levels of importance for the three scenarios by the job incumbents.

Table 7

Initial Item Pool: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 1--Combat

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	259	9.8	8.6	17.6	15.2	48.8	3.85	.33
64C	4	216	13.2	1.8	3.7	8.4	72.8	4.26	.74
71L	4	119	26.9	28.3	20.4	10.6	13.6	2.50	.74
95B	5	221	28.3	12.1	17.9	19.4	22.2	2.95	.88
Batch B									
11B	5	200	25.9	0.9	1.6	5.6	66.0	3.85	.73
19E	5	219	13.5	11.5	20.8	19.8	34.3	3.50	.66
31C	5	200	3.1	4.3	18.4	38.4	35.3	3.99	.53
63B	5	282	44.7	19.1	21.3	9.9	4.9	2.11	.69
91A	5	306	63.9	3.9	9.6	6.8	15.7	2.06	.74
Batch Z									
12B	6	229	38.1	12.7	13.3	20.9	14.9	2.62	.90
16S	5	208	11.7	3.8	11.5	11.3	61.6	4.07	.88
27E	5	233	13.9	7.6	24.0	36.8	17.6	3.36	.90
51B	4	228	0.8	22.2	30.4	16.2	30.4	3.53	.81
54E	7	162	15.9	2.9	13.5	18.6	49.0	3.82	.73
55B	5	236	0.8	3.4	60.7	30.4	4.6	3.34	.64
67N	6	221	14.5	12.7	29.0	23.5	20.2	3.22	.88
76W	5	214	6.5	13.3	14.0	13.1	53.1	3.93	.63
76Y	5	198	69.3	2.1	1.5	5.5	21.5	2.08	.90
94B	5	213	28.2	10.3	11.4	10.6	39.3	3.22	.64

19K	6	250	27.2	4.6	14.9	17.1	36.1	3.30	.92

Table 7 (Continued)

Initial Item Pool: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 2--Combat Readiness

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	259	9.2	9.4	17.5	19.3	44.8	3.81	.34
64C	4	216	0.0	0.6	4.0	9.1	86.2	4.81	.75
71L	4	119	12.7	25.4	31.9	22.4	7.5	2.87	.74
95B	5	221	13.7	18.0	27.2	28.2	12.8	3.08	.74
Batch B									
11B	5	200	26.0	2.2	1.0	7.0	62.9	3.79	.69
19E	5	219	13.2	11.6	21.1	20.2	33.9	3.50	.66
31C	5	200	2.7	3.4	21.8	48.9	23.2	3.86	.47
63B	5	282	26.9	23.0	19.9	25.8	4.3	2.57	.67
91A	5	306	51.6	6.6	13.2	10.9	17.6	2.36	.75
Batch Z									
12B	6	229	22.8	16.5	15.9	29.1	15.5	2.98	.89
16S	5	208	9.4	4.5	11.1	11.5	63.4	4.15	.87
27E	5	233	11.1	6.3	17.2	36.6	28.7	3.65	.90
51B	4	228	0.5	12.1	43.5	16.0	27.8	3.58	.52
54E	7	162	15.8	2.8	15.2	20.6	45.6	3.77	.70
55B	5	236	0.6	2.5	50.5	40.7	5.66	3.48	.73
67N	6	221	12.2	5.9	14.5	39.2	28.1	3.65	.92
76W	5	214	3.4	3.4	11.2	12.0	69.9	4.41	.34
76Y	5	198	59.2	4.9	5.9	7.1	22.8	2.29	.88
94B	5	213	24.1	13.5	17.8	9.6	35.0	3.18	.63

19K	6	250	26.9	5.0	15.3	20.3	32.5	3.26	.92

Table 7 (continued)

Initial Item Pool: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 3--Garrison Duty

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	259	10.2	10.3	17.2	17.5	45.0	3.77	.17
64C	4	216	0.0	0.6	1.9	5.8	91.6	4.88	.67
71L	4	119	7.0	14.7	33.1	36.0	9.1	3.25	.48
95B	5	221	31.7	10.4	15.0	19.5	23.4	2.92	.85
Batch B									
11B	5	200	28.4	2.9	4.1	5.6	59.0	3.69	.70
19E	5	219	8.8	9.8	21.9	22.3	37.2	3.69	.70
31C	5	200	2.6	2.9	24.0	53.6	16.9	3.79	.45
63B	5	282	13.7	13.3	26.0	37.6	9.33	3.15	.53
91A	5	306	33.1	7.6	17.5	20.3	21.5	2.90	.71
Batch Z									
12B	6	229	3.2	12.1	14.9	31.9	37.9	3.89	.81
16S	5	208	0.9	1.2	3.3	6.7	87.9	4.79	.75
27E	5	233	10.9	5.9	14.2	26.8	42.1	3.83	.89
51B	4	228	0.0	4.2	10.9	5.5	79.4	4.60	.00
54E	7	162	16.0	5.7	31.9	24.2	22.1	3.31	.72
55B	5	236	0.2	2.1	44.4	44.8	8.4	3.59	.70
67N	6	221	12.4	3.8	14.8	18.2	50.7	3.91	.91
76W	5	214	0.2	0.7	2.6	9.3	87.1	4.82	.26
76Y	5	198	1.9	2.1	16.2	13.8	65.9	4.40	.43
94B	5	213	18.0	23.6	12.8	6.9	39.6	3.25	.34

19K	6	250	25.5	7.2	23.3	16.7	27.2	3.13	.90

Table 8 contains the percentage of items rated Very Important (5) and Of Little Importance (1) by job incumbents under two sce. rios for Batches A, B, and Z. The mean of the mean importance scores across raters and all scenarios is also shown.

Table 8

Initial Item Pool: Mean Percent Importance Ratings and Mean Importance Score - Batches A, B, and Z

	Importance Rating (%)		Mean Importance Score
	Low (1)	High (5)	
Incumbents			3.52[a]
Combat Scenario	22.81	33.10	-
Garrison Scenario	11.24	43.06	-
Trainers	4.10	54.40	4.18

[a] All scenarios.

This table also contains item pool importance rating data for school trainers. Trainers did not rate items by scenarios; instead they rated how important it is for trainees to learn the knowledge represented by the item. Comparisons must be drawn between the incumbent means across scenarios and the trainers' means. More detailed information on trainer ratings will be presented in the section on review by trainers.

Two points about the data shown in Tables 6 and 7 are worth highlighting. First, the job incumbents rated a relatively large percentage of the items in the Very Important category. Using the combat scenario, they rated an average of 33.1% of the items Very Important; using the garrison duty scenario, they rated an average of 43.1% of the items as Very Important. Second, when importance ratings under the two scenarios are compared, the combat scenario appears to focus importance on fewer items. Thus, when the combat scenario is used, a lower percentage of items is rated as Very Important than when the garrison scenario is used (33.1 vs. 43.1%) and a higher percentage of items is considered to be Of Little Importance (22.8 vs. 11.2%). A 2x2 contingency table comparing item frequencies (Garrison & Combat vs. Rating 1 & 5) yields a chi square of 224.09, $\underline{p}$ = .004.

A possible explanation as to why job incumbents rated a lower percentage of items as Very Important when using the combat scenario is that incumbents focus their attention on a narrower set of activities in a combat setting. For example, correctly filling out forms will probably be unimportant in a combat scenario. The major concerns in combat would focus on activities that involve survival and control of the enemy. The major concerns in a garrison scenario, on the other hand, would include a wider range of activities.

Mean interrater reliabilities for the incumbents were reasonably high for the combat and combat readiness scenarios, .74 and .71 respectively, but significantly lower for the garrison scenario, .60 ($\underline{t}$ = 3.07, $\underline{p}$ = .006 and $\underline{t}$ = 2.96, $\underline{p}$ = .007). Overall interrater reliability was .67.

Relevance Ratings. The job relevance of draft test items was determined by asking incumbents, "Do Skill Level 1 personnel in this MOS need to use this knowledge on the job?" Since an MOS comprises many jobs, or duty positions, it seemed likely that incumbents in different billets might disagree about item relevance because they defined the job differently. The procedure followed was to favor inclusion: If any one respondent in the group asserted that the knowledge was required for job performance, then the item was flagged as job-relevant. The results of this procedure will be reported in the section on review by school trainers, which describes how relevance data were also obtained from trainers at MOS training sites.

REVIEW BY SCHOOL TRAINERS

Items in the initial item pool were also reviewed by school trainers at one of the training sites for each MOS. As with the review by job incumbents, the trainers reviewed the items for technical accuracy and appropriate vocabulary, and rated item content for importance and for relevance. It was during such site visits that the item trials were conducted with trainees, as described in the section on school tests.

Item Quality. The accuracy and appropriateness of the items were reviewed from the trainers' point of view, following essentially the same procedures described for job incumbents. Trainers were asked whether the item would be clear, whether distractors were realistic, and so forth. Items were then revised accordingly: unrealistic distractors replaced, stems modified, and so forth.

Importance Ratings. To obtain a measure of item importance from the trainers' point of view, each SME was given the following instructions:

> Look at each of the test questions and ask yourself how important it is that a trainee in the course learn the knowledge represented by this question.

Trainers used the same scale as incumbents to rate items' importance, but did not make use of different scenarios. Table 9 shows the percentage of items in the item pool that trainers rated at different importance levels for the various MOS. The table also contains interrater reliabilities for all MOS.

In general, trainers tended to rate items significantly higher than incumbents, as was shown in Table 7. Mean importance rating by trainers for the initial item pool was 4.18 (median = 4.03) while the mean of the means across scenarios for job incumbents on the item pool was 3.52 (median = 3.58)

Table 9

Initial Item Pool: Percentage of Items Rated at Five Importance Levels by School Trainers

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	6	297	3.9	2.8	12.7	10.5	70.1	4.40	.72
64C	7	215	0.4	2.1	12.6	42.1	42.7	4.24	.78
71L	5	122	22.1	5.7	6.4	3.6	62.1	3.78	.95
95B	5	122	0.5	0.5	3.3	17.9	77.8	4.72	.50
Batch B									
11B	7	200	6.9	8.1	18.5	26.9	39.5	3.83	.52
19E	6	214	4.5	7.8	17.3	23.0	47.3	4.01	.64
31C	6	192	3.6	3.6	33.8	38.0	20.8	3.69	.69
63B	6	238	5.6	13.1	40.5	23.7	17.1	3.34	.61
91A/B	6	299	3.8	8.7	21.3	28.9	37.2	3.87	.81
Batch Z									
12B	6	221	8.2	3.9	14.0	26.1	47.7	4.01	.87
16S	5	208	4.5	2.8	16.1	25.8	50.8	4.15	.61
27E	6	219	3.4	5.2	11.6	42.7	37.0	4.04	.73
51B	4	218	0.2	1.3	5.0	5.6	87.9	4.79	.57
54E	6	220	1.8	3.8	25.4	41.2	27.6	3.89	.66
55B	6	227	1.5	1.2	3.9	7.3	86.1	4.75	.32
67N	4	215	0.5	0.5	13.4	21.1	64.5	4.49	.18
76W	3	214	0.3	0.1	6.2	11.2	82.1	4.75	.00
76Y	6	132	0.0	0.1	0.9	0.2	98.7	4.98	.32
94B	6	200	6.2	9.4	19.9	28.2	36.3	3.79	.68

19K	6	202	2.5	1.1	22.8	31.9	41.7	4.09	.75

(Wilcoxon Z = 3.38, p - .001). This same trend appears in the proportions of items rated Very Important and Of Little Importance. On the average, trainers rated 54.4% of the items in the item pool as Very Important, while incumbents gave a Very Important rating to 33.1% of the items on the combat scenario and 43.1% of the items on the garrison scenario. Incumbents rated 22.8% of the items as being Of Little Importance on the combat scenario and 11.2% on the garrison scenario; trainers, however, rated only 4.1% of the items as Of Little Importance.

A possible explanation as to why the trainers rated a greater percentage of items as Very Important is that in a school setting, every piece of information with respect to the MOS is considered important. The school curriculum is designed to train and teach soldiers every MOS operation in a variety of military scenarios. Therefore, it makes sense that fewer items were rated as Very Important by job incumbents than by school trainers because the combat scenario is only a subset of what is taught in AIT.

Mean interrater reliability for the trainers across MOS was .58 (median = .62). This compares with a mean of .67 for incumbents (median = .70) across all three scenarios.

Relevance Ratings. To establish the relevance of the draft test items to training, trainers were asked the following:

> Can trainees be expected to have the knowledge
> represented in the items as a result of training?

As with job relevance, the procedure favored inclusion. If any of the trainers responded affirmatively, the item was flagged as training-relevant. At this point, relevance data were available for all items with respect to the job alone (from SME/Incumbents) and training alone (from SME/Trainers). Where the two judgments overlapped, items were considered relevant to both job and training.

Table 10 is based on relevance data obtained from job incumbents and from trainers and shows the distribution of the various classes of items for each MOS in the initial item pool, which formed the bases for the version of the test administered to trainees in the schools. The Not Rated category consists of items added to the pool after relevance ratings had been collected. Percentages were computed based on the summation of the Job-Only, Training-Only, and Job-and-Training categories.

As would be expected, many more items were rated as Job and Training (2,843 or 75.5%) than as either Job Only (676 or 17.9%) or Training Only (249 or 6.6%). Also, there are substantial differences in the range of items in these three categories. Of particular interest is the comparison between Job Only (range = 0-78) and Training Only (range = 0-140). The large range for Training Only is accounted for solely by MOS 91A; without this one MOS, the range would be 0-19. MOS 91A is the designation for medical specialists, and incumbents appear to believe that many items which trainers consider relevant are not relevant to the job.

Table 10

Initial Item Pool: Number and Percent of Items Rated Relevant to Job and Training

MOS	Job Only		Training Only		Job and Training		Not Relevant	Not Rated
	N	%	N	%	N	%	N	N
Batch A								
13B	70	41.4	5	3.0	94	55.6	6	62
64C	78	36.8	0	0.0	134	63.2	0	16
71L	42	34.4	4	3.3	76	62.3	0	0
95B	64	31.5	8	3.9	131	64.5	11	20
Batch B								
11B	68	39.5	14	8.1	90	52.3	21	25
19E	32	16.2	9	45.7	156	79.2	2	5
31C	47	26.3	15	8.4	117	65.4	5	8
63B	48	23.0	8	3.8	153	73.2	2	4
91A	0	0.0	140	54.9	115	45.1	5	5
Batch Z								
12B	7	3.4	0	0.0	197	96.6	0	23
16S	11	5.4	0	0.0	191	94.6	0	6
27E	1	0.5	19	9.3	185	90.2	0	15
51B	0	0.0	0	0.0	202	100.0	0	16
54E	0	0.0	1	0.5	207	99.5	0	15
55B	0	0.0	5	2.4	206	97.6	0	16
67N	1	0.5	0	0.0	208	99.5	0	8
76W	68	31.8	12	5.6	134	62.6	0	0
76Y	78	39.2	0	0.0	121	60.8	0	1
94B	61	31.1	9	4.6	26	64.3	8	2
Total	676	17.9	249	6.6	2843	75.5	60	277

Given the doctrinal emphasis on relating training to the job, it is not surprising that (with the exception of MOS 91A) not very many items were rated as Training Only--this despite the effort by item writers to create such items within their budgets.

SCHOOL TEST ADMINISTRATION

After review by job incumbents and trainers, test items were administered to groups of trainees in their last week of training. A sample of trainees was also interviewed after the tests to obtain information about item clarity and comprehensibility. Specific questions included the following:

- Did you have any difficulty understanding the question? Were there any words or phrases which were difficult to understand?
- Do you agree with the correct answer? Is there a better way to state the answer?
- (For items derived from tasks performed in training) Is it necessary to know the answer to this question to perform the task in training?
- (For items derived from tasks performed in training) Is the item a fair measure of a soldier's ability to perform the task?

The results of this test administration to trainees are shown in Table 11. All these results are based on items relevant to training, that is, job-and-training and training-only items. Items relevant only to the job are not included in these data.

When tests were administered in the schools, the targeted number of subjects was 50 at each school. The range of subjects to whom the tests were actually administered was from 32 for MOS 76W to 71 for MOS 16S; the mean was 50.1 subjects.

In general, the school test versions obtained high test reliabilities. Alpha of tests administered to trainees ranged from .789 to .972 with a mean reliability of .90 across all tests. The few school tests that obtained lower reliabilities (e.g., MOS 71L, alpha = .79) were the tests with fewer items (e.g., MOS 71L, N=71). It is reasonable to expect that the longer tests would generate higher reliabilities because a larger sample of test items is more likely to arrive at a more adequate and consistent measure. Lower reliabilities would be improved by lengthening tests.

An index of difficulty was computed by dividing the mean number of items correct by the number of items, that is, the percentage of items on a test that were correct. This percentage ranged from 41.4 for MOS 63B to 67.7 for MOS 55B. The mean percentage correct was 54.4.

PREPARATION FOR FIELD TEST OF BATCHES A AND B

After trainee tryouts at the schools were completed and items revised in accordance with the trainers' and trainees' comments, the item pools for the Batch A and Batch B MOS were prepared for field test administration to job

Table 11

Results from School Tests Administered to Trainees

MOS	Number of Subjects	Number of Items	Mean Number Correct	SD	Range	Alpha	Mean Percent Correct
Batch A							
13B	50	104	54.40	10.25	44	.81	52.3
64C	50	130	69.02	13.74	60	.87	53.1
71L	70	71	39.30	7.4	31	.79	55.3
95B	50	105	69.56	10.59	46	.85	66.2
Batch B							
11B	51	111	53.39	13.70	74	.91	48.1
19E	50	169	102.04	18.36	86	.92	60.4
31C	49	135	78.31	14.63	71	.90	58.0
63B	60	162	67.06	19.77	78	.92	41.4
91A	49	255	128.10	40.44	201	.97	50.2
Batch Z							
12B	50	214	118.06	16.56	78	.88	55.4
16S	71	197	120.01	18.98	112	.91	60.9
27E	43	219	131.28	21.54	102	.92	59.9
51B	50	218	120.46	21.97	107	.93	55.2
54E	46	220	131.15	19.76	75	.91	59.6
55B	48	227	153.63	21.59	101	.92	67.7
67N	47	214	122.55	19.91	108	.91	57.3
76W	32	146	67.13	15.15	58	.89	46.0
76Y	50	122	68.80	19.02	84	.94	56.1
94B	45	168	76.69	18.23	74	.90	45.6

incumbents. Data from the field test administration were later used (along with data from the administration of the school test to trainees, relevance data, and importance data) to convert the pools of draft items into the Job-Relevant Knowledge Tests to be used for Concurrent Validation.

As the pools were cut and items added or changed in preparation for field testing, the descriptive characteristics of the overall pools--that is importance and relevance--inevitably changed as well. The characteristics of the field test versions in terms of importance and relevance are reported below. These data parallel those reported for the initial item pools.

Importance Ratings. Table 12 shows, for the field test versions, the percentage of items that job incumbents had earlier rated at each of the five levels of importance for the three scenarios. Since the field tests included only Batches A & B, the data reported in Table 12 are for those nine MOS. Most of these tests had been culled of items in preparation for the field

Table 12

Field Test Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 1--Combat

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	180	8.5	7.9	18.3	15.5	29.8	3.90	.57
64C	4	210	13.2	1.8	3.7	8.4	72.8	4.26	.74
71L	4	119	27.0	28.3	20.4	10.6	13.6	2.55	.74
95B	5	221	28.3	12.1	17.9	19.4	22.2	2.95	.88
Batch B									
11B	5	143	25.4	0.8	1.4	5.9	66.4	3.87	.73
19E	5	202	11.3	11.7	21.3	19.8	35.9	3.57	.64
31C	5	187	2.1	4.6	18.5	38.3	36.4	4.02	.50
63B	5	216	43.7	18.8	25.1	12.0	6.2	2.15	.72
91A	5	260	64.7	3.8	9.6	6.5	15.4	2.04	.74

Table 12 (Continued)

Field Test Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 2--Combat Readiness

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	180	7.7	8.3	17.9	19.9	46.2	3.89	.58
64C	4	210	0.0	0.6	4.0	9.1	86.2	4.81	.75
71L	4	119	12.7	25.4	32.0	22.4	7.5	2.87	.74
95B	5	221	13.7	18.0	27.2	28.2	12.8	3.08	.74
Batch B									
11B	5	143	24.9	2.5	1.9	7.0	63.6	3.82	.71
19E	5	202	11.2	11.6	21.5	20.2	35.5	3.57	.63
31C	5	187	1.7	3.6	21.8	48.8	24.1	3.90	.45
63B	5	216	25.8	23.6	19.3	26.4	4.8	2.61	.64
91A	5	260	51.8	6.5	13.5	10.7	17.5	2.35	.74

Table 12 (Continued)

Field Test Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 3--Garrison Duty

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	180	9.3	8.0	17.2	18.1	46.5	3.83	.36
64C	4	210	0.0	0.6	1.9	5.8	91.6	4.88	.67
71L	4	119	7.0	14.7	33.1	36.0	9.1	3.25	.48
95B	5	221	31.7	10.4	15.0	19.5	23.4	2.92	.85
Batch B									
11B	5	143	26.0	3.5	3.8	5.0	61.7	3.73	.76
19E	5	202	7.1	9.7	21.6	22.6	39.0	3.77	.66
31C	5	187	1.6	3.1	23.9	53.6	17.7	3.83	.49
63B	5	216	11.7	13.7	26.1	37.9	10.6	3.22	.49
91A	5	260	31.9	7.8	18.0	20.5	21.7	2.92	.70

tests and consequently are shorter than tests in the item pool (Table 6). The basis for the culling has already been described in detail. In addition, prior to the field test some items were added to the item pool on which importance data had not been collected and for which no importance ratings were available.

As would be expected, the pattern of importance ratings by incumbents across scenarios was little affected by the culling procedure. The mean of mean importance ratings across MOS is lower for combat (mean = 3.26) than for combat readiness (mean = 3.43) and for garrison (mean = 3.59) scenarios.

When initial item pool and field test versions are compared (Table 13), there are small differences in the percentage of items incumbents rated Very Important and Of Little Importance on the combat scenario (Very Important: 34.8 to 34.3% and Of Little Importance: 25.5 to 24.9%) and the garrison scenario (Very Important: 34.8 to 35.7% and Of Little Importance: 15.1 to 14.0%). These changes are generally in the direction that would be expected, given the procedures that were used to cull the initial item pools. There was little difference in the mean across all scenarios between the item pool (3.40) and field test versions (3.43).

Table 13

Comparison of Field Test to Item Pool: Mean Percent Importance Ratings and Mean Importance Score - Batches A and B

	Initial Item Pool			Field Test		
	Importance Rating (%)		Mean Importance Score	Importance Rating (%)		Mean Importance Score
	Low (1)	High (5)		Low (1)	High (5)	
Incumbents			3.40			3.43
Combat Scenario	25.48	34.84	-	24.90	34.30	-
Garrison Scenario	15.06	34.78	-	14.03	35.70	-
Trainers	5.70	46.07	3.97	4.97	46.65	4.02

The mean interrater reliabilities of importance ratings were slightly lower on the field test version than on the item pool for the combat (mean = .69) and combat readiness scenarios (mean = .67). The interrater reliabilities on the garrison scenarios were identical (x = .60). On the average, the raters agreed on items' level of importance to the job.

Table 14 shows the average percentage of items in the field test which school trainers had rated at different importance levels. The table also contains interrater reliabilities for Batches A & B.

Table 14

Field Test Version: Percentage of Items Rated at Five Importance Levels by School Trainers

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	6	152	3.9	2.8	12.6	10.5	70.1	4.40	.73
64C	7	145	0.4	2.1	12.6	42.1	42.7	4.24	.78
71L	5	122	22.1	5.7	6.4	3.6	62.1	3.78	.95
95B	5	90	0.5	0.5	3.3	17.9	77.8	4.72	.50
Batch B									
11B	7	144	5.0	7.5	17.7	29.4	40.4	3.93	.41
19E	6	202	3.4	7.7	17.4	22.8	48.7	4.06	.62
31C	6	187	2.9	3.7	34.3	38.1	20.9	3.70	.62
63B	6	216	3.4	11.6	41.0	25.5	18.4	3.44	.48
91A	6	260	3.1	6.9	21.1	30.1	38.8	3.95	.78

As expected for the culled tests, mean importance ratings were somewhat higher for field tests than for the item pools for both trainers (mean = 4.02 vs. mean = 3.97) and incumbents (mean = 3.43 vs. mean = 3.40). As discussed earlier in connection with the initial item pool, trainers rated items higher overall than did incumbents (Table 13). Mean trainer interrater reliability across MOS was .53 (median = .59) which compared with a mean of .65 for incumbents (median = .68) across all three scenarios.

Relevance Ratings. Table 15 contains the relevance rating data for the version of the test administered to incumbents in the field tests. The distribution across relevance categories is similar to that noted in the earlier version used for school testing (see Table 10).

Table 15

Field Test Version: Number and Percent of Items Rated Relevant to Job and Training

MOS	Job Only		Training Only		Job and Training		Not Relevant	Not Rated
	N	%	N	%	N	%	N	N
Batch A								
13B	70	41.2	5	2.9	95	55.9	6	59
64C	80	37.2	0	0.0	135	62.8	0	13
71L	42	34.4	4	3.3	76	62.3	0	8
95B	64	31.5	8	3.9	131	64.5	11	20
Batch B								
11B	68	39.5	14	8.1	90	52.3	21	26
19E	32	16.2	9	4.6	156	79.2	2	5
31C	47	26.3	15	8.4	117	65.4	5	20
63B	48	23.0	8	3.8	153	73.2	2	8
91A	0	0.0	140	54.9	115	45.1	5	5
Total	451	26.2	203	11.8	1068	62.0	52	164

FIELD TEST WITH JOB INCUMBENTS

Procedure

Field testing was conducted in two phases--from March through September 1984 for the Batch A MOS, and from February through April 1985 for the Batch B MOS. In each MOS, incumbents were tested for 2 full days. The JRKT administration took four hours. The hands-on and knowledge task performance tests each required one half day of participant time; the predictor test battery required a 4-hour block; and the other 4-hour block was used for administration of various rating scales and questionnaires. These other measures are described in Pulakos and Borman (1986), and Campbell, Campbell, Rumsey, and Edwards (1986). The field test locations and numbers of soldiers tested in each location are shown in Table 16.

Table 16

Soldiers by MOS by Location of Field Test

	Location						
MOS	Fort Hood	Fort Lewis	Fort Polk	Fort Riley	Fort Stewart	USAREUR	Total
Batch A							
13B						150	150
64C						155	155
71L	48		60	21			129
95B	42		42	30			114
Batch B							
11B		29	30	30	31	58	178
19E		30	31	24	30	57	172
31C		16	26	26	23	57	148
63B		13	26	29	27	61	156
91A		24	30	34	21	58	167
Total	90	112	245	194	132	596	1369

At each site, an officer and two NCOs from one of the supporting units were assigned to support the field test. The officer provided liaison between the data collection team and the tested units, and the NCOs coordinated the acquisition of equipment and personnel. At each site a test site manager from the project staff supervised all research activity and maintained the orderly flow of personnel through the data collection points.

Before any instruments were administered, each soldier was asked to read a Privacy Act Statement, DA Form 4368-R. Project staff then gave a brief introduction on the purpose of the project, emphasizing the confidentiality of the data, and administered a Background Information Form. Soldiers moved in groups of about 15 to either the hands-on testing, one of the knowledge test sessions, or a rating session. The order of administration of the measures was counterbalanced across groups and locations within MOS.

After soldiers appeared for testing, their first- and second-line supervisors were identified and notified of the scheduled supervisor rating session. Considerable flexibility was necessary in providing alternate sessions for supervisors, including offering evening and weekend times for individuals. Each supervisor session normally took 2 to 3 hours.

Project staff members served as the test administrators for the JRKT. Times to complete each test booklet were recorded to assist in reducing the 4-hour block for the field test to the 2-hour block for the Concurrent Validation. Instructions for administering the tests are shown in Appendix A in ARI Research Note in preparation.

The JRKT were grouped into three booklets containing about equal numbers of items. Each booklet required about 50 minutes to complete, with a 10-15 minute break between booklets. The order of the booklets was counterbalanced among the soldiers in each group. The purpose for dividing the material into separate booklets was to try to control the effects of fatigue and waning interest.

Results of Field Testing

The results of the administration of the field tests to incumbents in the MOS in Batches A and B are shown in Table 17. Test scores are based on items relevant to the job, that is, job-and-training and job-only items. Items relevant only to training are not included in the results shown.

Table 17

Results from Field Tests Administered to Incumbents

MOS	Number of Subjects	Number of Items	Mean Number Correct	SD	Range	Alpha	Mean Percent Correct
Batch A							
13B	149	133	49.19	16.47	74	.90	44.5
64C	155	137	70.32	17.23	75	.91	51.3
71L	129	97	50.54	9.94	51	.83	52.1
95B	112	131	77.29	10.18	51	.76	59.0
Batch B							
11B	166	162	86.42	19.99	98	.93	53.3
19E	169	193	112.89	20.98	142	.93	58.5
31C	143	176	99.63	20.14	120	.92	55.6
63B	155	205	106.92	19.38	107	.90	52.1
91A	155	115	72.95	10.29	76	.82	63.4

The number of job incumbents to whom the tests were administered ranged from 112 for MOS 95B to 169 for MOS 19E. The mean number of subjects was 148.1. Item statistics (i.e., biserial correlations and proportion correct) were computed for all test items. All of the tests have relatively high reliability coefficients. Alpha of tests administered to job incumbents ranged from .76 for MOS 95B to .93 for MOS 19E with a mean reliability across all nine tests of .88. The percentage correct for job incumbents ranged from 44.5 for MOS 13B to 63.4 for MOS 91A with a mean correct of 54.5%.

The equivalent figures reported for the earlier administration to trainees were for all 19 MOS. When these trainee figures are recomputed for only the nine MOS that participated in the field tests, results for trainees and for field test job incumbents match closely. Mean trainee alpha was .88, and mean incumbent alpha was .88. Mean correct for trainees was 53.9%, compared to 54.5% for incumbents.

One MOS of particular interest is 91A where there was a substantial drop in the number of test items on the test as a whole (field test version = 260) versus a subset of items (job only and job-and-training = 115). This drop is accounted for by the fact, as noted in the discussion of Table 10, that many 91A items were rated as relevant to training only.

REVIEW BY TRADOC PROPONENT AGENCIES

All pre-Concurrent Validation JRKT versions were submitted to the appropriate Army Training and Doctrine Command Proponent for review. The number of items sent out for review and the number of items cut, added, or modified as a result of review are summarized in Tables 18, 19, and 20. These tables, which also show the number of items dropped from the pools on the basis of nonrelevance, low importance, or item characteristics, are discussed in the following section.

PREPARATION FOR CONCURRENT VALIDATION TEST

Procedure for Reducing Test Length

It was generally agreed that a suitable test length for the Concurrent Validation (Batch A and B MOS) would be about 150 items. (This number was an approximation based on the 2-hour period available for Task 3 JRKT testing and data regarding the number of minutes per item soldiers needed to complete the Batch A tests.)

To reduce the size of the item pools as required, any item that had been rated not relevant to the job and also not relevant to training was dropped first. To reduce test length further where needed, items were dropped that were lowest in importance and/or highest in difficulty. Because the performance domain was assumed to be multidimensional, items were not generally eliminated solely on the basis of a negative biserial correlation with the rest of the test. However, some items were dropped that exhibited the three characteristics of (a) low pass rate, (b) negative biserial, and (c) a distractor or distractors with a high positive biserial.

Tables 18, 19, and 20 report for Batches A, B, and Z, respectively, the number of items remaining on the tests after all cuts had been made. Versions of the tests used for the Concurrent Validation will contain the number of items shown in the columns on the far right. The tables for Batches A and B differ slightly from the table for Batch Z. Many of the Batch A and B cuts (Tables 18 and 19) were made using field test data, which did not exist for Batch Z, as noted previously. Therefore, Table 20, reporting Batch Z data, begins with Number of Items sent to the Proponent (Column 2).

During the cutting of the item pools, an effort was made to keep the relative frequency of items in each AOSP duty area about the same as it had been before the Proponent review and, in particular, to avoid inadvertently eliminating any duty area. To maintain the intended balance of coverage over duty areas, items were added back to, as well as deleted from, the pool. Figure 4 shows how a simple spreadsheet program was used in reducing the total number of items in MOS 67N from 201 to 175, without causing any one duty area to gain or lose more than 20% of its previous share of the test.

Finally, to allow measuring examinees' loss of motivation during the testing period, five low-difficulty items were moved to the beginning of each test, and five to the end. Comparing performance on the two sets during the Concurrent Validation may reveal evidence of guessing or signs of test fatigue. The placing of five easy items at the beginning of the tests was also intended to be a motivating factor in itself.

The tests differ greatly in type of content and total coverage, and therefore their length varies. Another factor that influenced the length of the tests was the fact that Batch Z had not been field tested. Some item analysis data for Batch Z were available, but only for trainees. An analysis of the performance of incumbents and trainees on Batch A and Batch B tests, for whom data were available, suggested that it would be unwise to make cuts in Batch Z JRKT using trainee data. In the absence of complete item analysis data from both trainees and incumbents, these cuts were made on the basis of item importance ratings (items of lower importance were dropped).

Once item analysis data become available for Batch Z trainees and incumbents--that is, after the Concurrent Validation--the tests can be cut on that basis. More than 150 items were left in the Batch Z tests so that the tests could be cut to about 150 items on the basis of additional data.

Characteristics of Concurrent Validation Version of the Tests

Table 21 shows the percentage of items in the Concurrent Validation versions of the tests that job incumbents had rated at each of the five levels of importance for the three scenarios. These tests had been further culled of items and are consequently shorter than tests in the field test versions. As would be expected, the pattern of importance ratings across scenarios was little affected by the culling procedure. The mean of mean importance ratings across MOS is lower for combat (mean = 3.29) than for combat readiness (mean = 3.51) and for garrison (mean = 3.88) scenarios.

Table 18

Number of Items in Tests at Each Stage of Development: Batch A

MOS	Initial Item Pool No.[a]	No. of Cuts by Category		No. of Items Sent to Proponent Review[b]	Proponent Review		No. of Items Remaining
		Not Relevant	Low Importance or Poor Item Characteristics		Cut/Added	Modified	
13B (SP)	163/68[c]	18	75	138	2/0	0	136
13B (T)	163/67[d]	14	55	161	5/0	0	156
64C	228	2	86	140	12/0	70	128
71L	130	1	28	101	6/10	12	105
95B	223	11	72	140	6/5	9	139

[a] Items field tested.

[b] Reflects one or more Proponent reviews.

[c] There were 163 items common between the SP & T versions and 68 items unique to the SP version.

[d] There were 163 items common between the SP & T versions and 67 items unique to the T version.

Table 19

Number of Items in Tests at Each Stage of Development: Batch B

	Item Pool No.		No. of Cuts by Category			Proponent Review		Items Added	
				Low Importance	No. of Items			to	
		Batch B		or Poor Item	Sent to			Rebalance	No. of Items
MOS	School	Field Test	Not Relevant	Characteristics	Proponent Review	Cut/Added	Modified	Budget	Remaining
11B	200	199	5	13	181	35/0	13	4	150
19E	214	202	2	21	179	17/0	8	0	162
31C	192	199	5	15	179	4/0	7	0	175
63B	238	217	2	47	168	24/23[a]	81	5	172
91A	299	260	5	46	209	34/0[b]	17	0	175

[a] Reflects two Proponent Reviews.

[b] Reflects an additional cut made in conjunction with Proponent review in order to bring the test down to 175 items.

Table 20

Number of Items in Tests at Each Stage of Development: Batch Z

MOS	No. of Items Sent to Proponent Review	Proponent Review		Additional Cuts Based on Low Importance or Poor Item Characteristics	Items Added to Rebalance Budget	No. of Items Remaining
		Cut/Added	Modified			
12B	211	0/0	5	35	0	176
16S	202	49/0	35	5	1	149
27E	205	2/0	9	28	0	175
51B	202	8/0	5	31	0	163[a]
54E	207	63/0	22	5	6	145
55B	212	0/0	5	31	0	181
67N	207	6/0	0	26	0	175
76W	195	1/0	0	19	0	175
76Y	188	2/0	11	11	0	175
94B	187	3/0	4	8	0	175

[a] Reduced to this number due to time-consuming math problems.

MOS: 67N

AREA	Original Number	Original Percent	Pre-PR Number	Pre-PR Percent	Orig-Pre Change	Post-PR Number	Post-PR Percent	Pre-Post Change	Post-PR Number	Post-PR Percent	Pre-Post Change
A	23	9.1	18	8.7	-0.05	17	8.5	-0.03	16	9.1	0.05
B	2	0.8	2	1.0	0.22	2	1.0	0.03	2	1.1	0.18
C		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
D	14	5.6	11	5.3	-0.04	11	5.5	0.03	11	6.3	0.18
E	25	9.9	22	10.6	0.07	22	10.9	0.03	17	9.7	-0.09
F	39	15.5	30	14.5	-0.06	29	14.4	.00	23	13.1	-0.09
G	20	7.9	17	8.2	0.03	17	8.5	0.03	16	9.1	0.11
H	20	7.9	17	8.2	0.03	16	8.0	-0.03	15	8.6	0.04
I	14	5.6	14	6.8	0.22	13	6.4	-0.04	13	7.4	0.10
J	6	2.4	4	1.9	-0.19	4	2.0	0.03	4	2.3	0.18
K	15	6.0	13	6.3	0.06	12	6.0	-0.05	12	6.9	0.09
L		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
M		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
N	10	4.0	7	3.4	-0.15	7	3.5	0.03	7	4.0	0.18
O		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
P		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
Q	50	19.8	44	21.3	0.07	44	21.9	0.03	32	18.3	-0.14
R		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
S		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
T		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
U		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
V		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
W		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
X		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
Y		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
Z		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
POI	14	5.6	8	3.9	-0.30	7	3.5	-0.10	7	4.0	0.04
LEAVE BLANK		0.0		0.0	0.00		0.0	0.00	0	0.0	0.00
	252	100.0	207	100.0		201	100.0		175	100.0	

Figure 4. Example of spreadsheet for adjusting item pool.

Table 21

Concurrent Validation Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios for Job Incumbents

Scenario 1--Combat

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	101	7.5	7.5	17.8	14.6	52.6	3.97	.51
64C	4	99	10.7	0.9	1.9	8.2	78.2	4.42	.71
71L	4	83	26.7	32.4	19.0	10.1	11.7	2.48	.71
95B	5	128	27.3	12.9	19.4	19.4	20.9	2.93	.89
Batch B									
11B	5	143	25.4	0.8	1.4	5.9	66.4	3.87	.73
19E	5	153	9.7	11.9	21.5	19.5	37.4	3.63	.59
31C	5	163	1.7	4.0	17.3	28.4	38.5	4.08	.41
63B	5	138	39.1	19.6	23.9	11.0	6.4	2.26	.69
91A	5	175	63.7	3.7	10.4	6.0	16.2	2.07	.75
Batch Z									
12B	6	162	34.4	12.2	12.9	23.3	17.2	2.77	.90
16S	5	148	11.2	3.8	11.7	11.6	61.6	4.09	.87
27E	5	172	5.5	7.3	24.5	41.6	21.0	3.65	.85
51B	4	146	0.7	22.1	30.3	15.7	31.2	3.55	.82
54E	7	100	10.4	2.6	14.8	19.8	52.3	4.01	.72
55B	5	178	0.8	3.0	60.0	31.0	5.2	3.37	.65
67N	6	169	9.4	13.9	32.3	23.6	20.8	3.32	.84
76W	5	170	6.6	13.5	14.6	12.9	52.3	3.91	.68
76Y	5	165	72.6	1.9	1.4	4.7	19.3	1.96	.90
94B	5	109	30.8	10.6	11.7	9.3	37.5	3.12	.61

Table 21 (Continued)

Concurrent Validation Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 2--Combat Readiness

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	101	6.2	7.8	17.4	19.7	48.9	3.97	.48
64C	4	99	0.0	0.6	2.6	7.6	89.2	4.85	.71
71L	4	83	10.5	28.9	33.0	20.9	6.7	2.84	.72
95B	5	128	9.5	18.7	29.7	30.3	11.8	3.16	.69
Batch B									
11B	5	143	24.9	2.5	1.9	7.0	63.6	3.82	.71
19E	5	153	9.7	11.9	21.7	19.7	37.0	3.62	.58
31C	5	163	1.3	3.3	19.7	50.2	25.4	3.95	.35
63B	5	138	23.5	21.7	20.1	28.4	6.2	2.72	.61
91A	5	175	47.8	7.0	14.6	11.2	19.4	2.47	.72
Batch Z									
12B	6	162	19.2	15.7	15.3	32.3	17.4	3.13	.88
16S	5	148	8.5	4.7	11.6	11.6	63.5	4.17	.87
27E	5	172	3.9	3.9	17.7	40.5	33.9	3.96	.84
51B	4	146	0.5	11.6	43.0	16.9	27.9	3.60	.50
54E	7	100	10.1	2.3	16.8	21.4	49.3	3.97	.66
55B	5	178	0.7	1.9	50.4	40.7	6.3	3.50	.74
67N	6	169	7.4	5.6	15.6	43.2	28.2	3.79	.87
76W	5	170	3.6	3.6	10.8	11.6	70.2	4.41	.39
76Y	5	165	61.9	4.7	6.3	6.8	20.2	2.19	.87
94B	5	109	27.8	14.4	15.7	9.7	32.3	3.04	.55

Table 21 (Continued)

Concurrent Validation Version: Percentage of Items Rated at Five Importance Levels for Three Scenarios by Job Incumbents

Scenario 3--Garrison Duty

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	7	101	8.3	8.3	17.4	17.8	48.1	3.89	.23
64C	4	99	0.0	0.3	1.6	5.1	93.0	4.91	.57
71L	4	83	4.8	16.2	34.6	35.5	8.0	3.28	.34
95B	5	128	26.8	6.9	14.6	22.1	29.5	3.21	.87
Batch B									
11B	5	143	26.0	3.5	3.8	5.0	61.7	3.73	.69
19E	5	153	5.2	9.9	22.0	22.3	40.5	3.83	.62
31C	5	163	1.5	2.9	22.1	54.6	18.9	3.86	.32
63B	5	138	9.6	13.2	25.6	38.1	13.5	3.33	.48
91A	5	175	28.1	7.3	17.5	21.7	25.4	3.09	.73
Batch Z									
12B	6	162	2.5	10.0	13.1	33.5	40.9	4.00	.79
16S	5	148	0.4	0.9	3.1	6.7	88.8	4.82	.76
27E	5	172	3.9	3.8	13.8	28.6	49.8	4.16	.81
51B	4	146	0.0	3.6	11.8	4.8	79.8	4.61	.00
54E	7	100	11.0	5.7	32.8	24.7	25.7	3.48	.75
55B	5	178	0.2	1.7	43.9	44.6	9.5	3.61	.70
67N	6	169	7.5	3.9	13.3	20.2	55.0	4.11	.85
76W	5	170	0.2	0.9	2.0	8.9	87.9	4.83	.26
76Y	5	165	2.1	2.1	15.4	13.1	67.4	4.42	.41
94B	5	109	19.1	23.4	12.7	4.8	29.9	3.23	.39

When initial item pool and Concurrent Validation versions are compared (Table 22), there is a small increase in the percentage of items rated Very Important and a small decrease in the proportion rated Of Little Importance on both the combat scenario (Very Important: 33.1 to 34.0% and Of Little Importance: 22.8 to 20.6%) and the garrison scenario (Very Important: 43.1 to 46.5% and Of Little Importance: 11.2 to 8.3%). These changes are all in the direction that would be expected, given the procedures that were used to cull the initial item pools.

Table 22

Comparison of Concurrent Validation Test to Item Pool: Mean Percent of Importance Ratings (1 and 5) by Job Incumbents

	Importance Rating (%)			
	Initial Item Pool		Concurrent Validation Test	
Scenario	Low (1)	High (5)	Low (1)	High (5)
Combat	22.81	33.10	20.64	34.04
Garrison	11.24	43.06	8.27	46.54

Mean importance ratings across MOS for item pool and Concurrent Validation versions of the tests for each scenario were also compared. All were in the expected direction (i.e., higher importance on the Concurrent Validation test version than the item pool), and two were significant when compared using the Wilcoxon Matched Pairs test: combat scenario (initial item pool versus Concurrent Validation version) $\underline{Z}$ = 1.73, $\underline{p}$ = .08; combat readiness (initial item pool versus Concurrent Validation version) $\underline{Z}$ = 2.01, $\underline{p}$ = .04; garrison scenario $\underline{Z}$ = 2.86, $\underline{p}$ = .004.

The pattern of mean interrater reliabilities was similar to that of the initial item pool but somewhat lower: combat scenario mean = .73, combat readiness mean = .67, and garrison mean = .56.

Table 23 shows the average percentage of items on the Concurrent Validation versions of the tests rated at different importance levels by trainers. Again we note that trainers tended to rate items higher than incumbents. As would be expected for the culled version of the test to be used in the Concurrent Validation, mean importance ratings were somewhat higher than for the item pool (incumbents, 3.56 vs. 3.52; trainers, 4.26 vs. 4.18).

Mean trainer interrater reliability across MOS was .53, which compared with a mean of .65 for incumbents across all three scenarios.

Table 23

Concurrent Validation Version: Percentage of Items Rated at Five Importance Levels by Trainers

	Number		Rating (%)						
MOS	Raters	Items	1-Of Little Importance	2-Somewhat Important	3-Moderately Important	4-Quite Important	5-Very Important	Mean Rating	Interrater Reliability
Batch A									
13B	6	172	3.6	1.9	10.4	10.6	73.5	4.48	.73
64C	7	99	0.0	1.0	10.4	42.2	46.3	4.34	.72
71L	5	86	12.5	3.8	5.8	4.3	73.5	4.23	.91
95B	5	85	0.3	0.6	2.93	18.3	78.0	4.73	.34
Batch B									
11B	7	144	5.0	7.5	17.7	29.4	40.3	3.93	.49
19E	6	153	3.0	6.9	17.6	22.6	49.8	4.09	.63
31C	6	163	2.4	2.9	33.2	39.0	22.4	3.76	.61
63B	6	138	1.9	8.9	39.4	28.1	21.6	3.58	.28
91A	6	175	2.6	3.5	16.1	30.1	47.7	4.17	.79
Batch Z									
12B	6	164	6.4	3.7	11.2	25.9	52.7	4.15	.88
16S	5	148	3.0	1.3	13.6	25.5	56.5	4.31	.58
27E	6	174	0.9	3.2	10.8	45.4	39.7	4.20	.66
51B	4	149	0.3	0.8	3.7	5.0	90.1	4.84	.48
54E	6	131	2.1	3.8	24.7	42.1	28.2	3.92	.66
55B	6	181	0.5	0.9	4.2	6.4	87.8	4.80	.28
67N	4	173	0.0	0.1	12.6	21.2	66.0	4.53	.00
76W	3	170	0.4	0.2	6.5	9.6	83.3	4.75	.00
76Y	6	117	0.0	0.1	1.0	0.3	98.6	4.97	.33
94B	6	115	6.7	9.8	19.4	27.8	36.2	2.77	.69

Table 24 contains the relevance data for the version of the tests to be administered as part of the Concurrent Validation. The distribution across relevance categories is nearly the same as the original item pool and much more similar to the item pool (Table 10) than to the field test version (Table 15), since Table 10 and Table 24 are based on 19 MOS and Table 15 is based on 9 MOS.

Appendix B provides the complete collection of the JRKTs prepared for use in the Concurrent Validation. The versions of the JRKTs used in the field tests are available from the Army Research Institute.

Table 24

Concurrent Validation Version: Number and Percent of Items Rated Relevant to Job and Training

MOS	Job Only		Training Only		Job and Training		Not Relevant	Not Rated
	N	%	N	%	N	%	N	N
Batch A								
13B	40	25.0	1	0.6	119	74.4	0	30
64C	5	4.7	1	0.9	101	94.4	0	0
71L	18	20.9	2	2.3	66	76.7	0	7
95B	30	24.2	6	4.8	88	71.0	0	5
Batch B								
11B	48	36.1	9	6.8	76	57.1	13	3
19E	24	15.4	5	3.2	127	81.4	1	3
31C	43	27.6	10	6.4	103	66.0	5	19
63B	31	22.3	4	2.0	104	74.8	0	0
91A	0	0.0	82	47.7	90	52.3	3	3
Batch Z								
12B	0	0.0	0	0.0	162	100.0	0	0
16S	1	0.7	0	0.0	142	99.3	0	0
27E	0	0.0	14	8.0	161	92.0	0	0
51B	1	0.6	3	1.9	152	97.4	0	0
54E	0	0.0	0	0.0	135	100.0	0	0
55B	0	0.0	4	2.2	176	97.8	0	0
67N	0	0.0	0	0.0	173	100.0	0	0
76W	47	27.6	8	4.7	115	67.6	0	0
76Y	55	33.1	0	0.0	111	66.9	0	0
94B	30	23.3	7	5.4	92	1.3		30
Total	373	13.2	156	5.5	2293	81.3	25	100

Chapter 4

LESSONS LEARNED

ITEM TRACKING

Developing more than 200 test items for each of 20 different MOS required keeping track of data on more than 4,000 test items, through several revisions for each MOS. An item summary sheet (ISS) was devised for each MOS item pool (e.g., Figure 5). The ISS contained the following information for each item: (1) a master number; (2) an AOSP reference; (3) a POI reference; (4) class; (5) a school test version number; (6) school test revisions; (7) a Proponent review number; (8) Proponent review revisions; and (9) a Concurrent Validation number.

MOS 12B

1	2	3	4	5	6	7	8	9
12BM001	I-4	B-2a	JT	001		001		001
12BM002	I-4	B-2a	JT	002		002		002
12BM003	I-12	B-2a	JT	003		003		003
12BM004	I-12	B-2a	JT	004				
12BM005	I-12	B-2a	JT	005	M&T	005	DR	

LEGEND

Column

1 Item Master Number
2 AOSP Reference
3 POI Reference
4 Class
5 Item No. for School Test
6 Revisions
7 Item No. After Proponent review
8 Changes Made by Proponent
9 Concurrent Validation No.

Figure 5. Example of item summary sheet.

This method of tracking items "by hand" evolved as Task 3 personnel gradually came to understand the magnitude of the bookkeeping problem. The method became very cumbersome, as the number of cells in the set of tables grew to well over 20,000 (the number of MOS times the number of items for each, times the number of revisions). Accordingly, some type of automated database program running on a small computer appears virtually necessary in an effort of this magnitude.

Tracking items is further complicated by the fact that, as items are reviewed, many are changed significantly. Judgments regarding relevance and importance refer, of course, to a particular item at a given point in time. After each item change, a judgment must be made as to whether or not the item is still the "same" item. If it is not, then the original item must be recorded as dropped, and a new item with a new master number entered at the end of the item pool.

An ideal tracking system, whether automated or manual, would include the following elements:

(1) The capability to associate a unique identifier (such as an 8-digit alphanumeric code) with each test item, without showing that identifier on versions of the test where it would be distracting to examinees.

(2) The capability quickly to renumber items in a version of the pool after a subset of items has been dropped or items have been rearranged.

(3) In a manual system, a built-in error checking procedure that does *not depend on inspection of item content.* If an automated database program were used, this requirement would presumably be unnecessary, as long as the data entry procedures were designed to prevent severing the association between item content and item identifier.

(4) Computer printouts, such as item analyses, should clearly identify the version of the test being analyzed. Item analyses should also include the full text of test items.

EVOLUTION OF ITEM BUDGETS

Budgets were originally developed, as noted above, to help assure that the content domain will be clear, representative, and relevant. The budgets also serve the important function of guiding and providing discipline to item writers who often do not understand the psychometric issues involved in test construction.

There appears to be some tendency to see the original budgets as fixed or "set in concrete," when in fact they are evolving. Working with subject matter specialists, test item writers inevitably discovered that there are tasks that are no longer performed or there are new tasks or new ways of doing old tasks. Since the original pool of items was larger than needed for the tests, it was possible to keep reworking the budgets, dropping items here and adding new ones there to ensure that the content domain was appropriately sampled. The important point to note is that the original budgets were a starting point and that those original budgets changed as items went through various reviews. Although the budgets for the tests used in the Concurrent Validation generally looked very much like the initial budgets, there were places in which they were quite different.

The problem of tracking budget changes and adding or eliminating items to maintain adequate and appropriate coverage is not of the same magnitude as the problem of following the course of individual item changes, but it is certainly complicated and time-consuming. The easiest way to track budgets is to set up spread sheets that forecast the number of items needed to cover specific content areas as the item pools evolve into actual tests that are administered in the field.

ITEM ANALYSES: EMPHASIS ON STATISTICAL INFORMATION

In a typical test construction effort, the individual who reviews knowledge test items with the help of an item analysis is a subject matter expert. Furthermore, he/she is generally concerned with 1 test, not 19. In order to create 19 JRKTs, following the complicated, time-consuming, and systematic procedures outlined above within the time frame allowed and within budgets, the various tasks were divided among individuals with different types of skills. Item writers, for example, were generally not psychometricians. Personnel who administered tests to trainees in a given MOS were not always the same individuals who conducted the earlier item reviews with SMEs. In brief, test builders were seldom fully informed about every facet of an MOS.

Under such circumstances, psychometricians tend to view item analyses more from a statistical perspective than a content perspective. In some cases, a person who has not been immersed in the content of an MOS can develop hypotheses about content and its impact on item statistics, but these hypotheses are speculative. The best solution to this problem is probably to use psychometricians for the entire development process, giving selected individuals full responsibility for developing all aspects of one or two MOS. This solution involves a tradeoff of time and skilled manpower. One could hire a large number of specialists to do the job in the time allowed or greatly increase the time. Either way the cost would significantly increase. The penalty is clear: items tend to be dropped or added for statistical reasons, rather than modified for content reasons. Many potentially good items are discarded, and some marginal items probably survive.

Chapter 5

SUMMATION

The major objective of Task 3 was to create content-valid and reliable Job-Relevant Knowledge Tests for measuring the cognitive component of training success. How successful has Task 3 been in efforts to achieve test objectives?

Before an attempt to answer this question, an important caveat should be discussed. As has been noted, this report deals primarily with Batches A and B, not Batch Z. We have included a discussion of preparation work on Batch Z as a matter of record and in order to round out the description of Task 3 activities through the end of the 1985 fiscal year. However, at the time of this report, Batch Z had not been field tested. Batch Z tests, which will be used in the Concurrent Validation, contain many items that would undoubtedly have been removed had data regarding incumbent performance been available beforehand. At this time Batch Z is made up of a pool of items that must be cut into tests, using item analysis data from school administrations and the Concurrent Validation. To the extent that the same procedures were used to develop all three Batches, the comments below apply, but the discussion is focused on Batches A and B.

The tests in Batches A and B may be evaluated from three perspectives. First, since content validity is so crucial in the evaluation of instruments designed to measure training success, one can examine the process by which the tests were developed and use some of the standards identified by Guion (1977) and others as criteria for evaluating that process. Second, one can consider the development process up to the point of the final Proponent review, which indeed was an added step in the process, and compare the tests before and after Proponent review. The assumption here is that if the tests undergo relatively little change (particularly fundamental change such as cutting items and/or adding new items) as a result of the final Proponent review, the development process as originally conceived was valid. Finally, one can look at more traditional measures, such as the reliability of the tests.

The developmental process did conform to the three criteria of domain clarity, content representativeness, and content relevance.

First, the domain was operationally identified and items were drawn from that domain. The developmental model prescribed that the initial items would be drawn from published Army literature. It was recognized from the start, however, that the published literature inevitably lags behind practice (i.e., doctrine and equipment). Therefore, some change was inevitable as subject matter experts examined items. Nevertheless, the changes were in most cases not dramatic; many concerned terminology or phrasing rather than content. Despite the weaknesses in the procedures used to collect these data, there is still substantial agreement suggesting that both test developers and subject matter experts independently developed/assigned items using a common overlapping referent.

With respect to the second criterion, content representativeness, the proportions of items assigned to different duty areas on different versions of the test--that is, from initial item pools to the final Proponent review--are similar (see, for example, Figure 4). Inevitably, there were changes in the percentage of items in any given duty area, but radical changes in the distribution of items across duty areas were not required. In a few cases, it was found that some duty areas were no longer performed as a part of an MOS or an MOS has been given some new responsibility, but changes of this magnitude were rare and were almost never in a major duty area that had many items allocated to it.

With respect to the third criterion, content relevance, the elaborate procedure by which Task 3 staff determined relevance has been described. Items judged as being not relevant to training and/or the job were eliminated. Moreover, relevance was judged in terms of importance. Only those items judged to be very important on one or more of the three scenarios were retained.

Finally, Guion has stressed the issue of fairness, a criterion not mentioned in our earlier discussion of content validity. To meet the standard of fairness, every effort was made, when items were reviewed by subject matter experts, to ensure that the review groups were balanced for race and gender. The data on this issue are reported in Table 5.

Next we turn to the question of the extent to which the Proponent review altered or changed the JRKT tests. The short answer to this question is that, with one or two exceptions, not very many significant changes were made. Proponents requested three types of changes: cuts, additions, and modifications (Tables 18, 19, and 20). The mean percentages of these changes across all 19 MOS were as follows: cuts, 7.5%; additions, 1.4%; and modifications, 9.4%. When one considers the lengths of the tests, these percentages are not very great. Furthermore, modifications were in many cases relatively trivial and did not concern content so much as format or phrasing. The distributions of these changes were, however, quite skewed; certainly, in 3 or 4 cases out of the total possible number of 57 (three types of change X 19 MOS) they were unusually large, suggesting substantial disagreement. By consulting Tables 18, 19, and 20, one can note that the most significant disagreements occurred for MOS 16S (cuts), 54E (cuts), 11P (cuts), and 63B (modifications).

Finally, the tests can be evaluated in terms of more traditional psychometric measurements, particularly reliability. All of the tests have relatively high reliability coefficients. Academic batteries commonly have high reliability coefficients, generally ranging from .66 to .98, and all of the tests in Batches A and B approach the median level (.92).

Based on the data presented, one can conclude that the JRKT versions that were developed are reliable and content-valid measures of the cognitive component of training success. The item content of the tests was meticulously determined by item budgets. The actual items were written based on pertinent references (e.g., AOSP, POI). Furthermore, all items were evaluated by job incumbents, school trainers, and the respective Proponents. The tests were administered to actual school trainees. The test and item parameters were carefully analyzed. Based on the available data, each JRKT was carefully tailored to ensure that the test content was a reliable and valid representation of training success.

REFERENCES

Campbell, Charlotte C., Campbell, Roy C., Rumsey, Michael G., & Edwards, Dorothy C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: Army Research Institute. In preparation.

Campbell, John P. (Ed.). (1987). Improving the selection, classification, utilization of Army enlisted personnel: Annual report, 1985 fiscal year (ARI Technical Report 746). In preparation.

Eaton, Newell K., & Goer, Marvin H. (Eds.). (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report (ARI Research Note 83-37). (ADA 137 117)

Eaton, Newell K., Goer, Marvin H., Harris, James H., & Zook, Lola M. (Eds.). (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (ARI Technical Report 660). (ADA 178 944)

Guion, Robert M. (1977). Content validity--The source of my discontent. Applied Psychology Measurement, 1 (winter), 1-10.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report (ARI Research Report 1347). (ADA 141 807)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Project A - Research plan (ARI Research Report 1332). (ADA 129 728)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report synopsis, 1984 fiscal year (ARI Research Report 1393). (ADA 173 824)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Appendices to annual report, 1984 fiscal year (ARI Research Note 85-14).

Pulakos, Elaine D., & Borman, Walter C. (Eds.). (1986). Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). Alexandria, VA: Army Research Institute. (ADB 112 857)